National Records of Scotland

PAMS 19 (07)

Population And Migration Statistics (PAMS) Committee (Scotland)

Update on 2001-2010 SAPE reestimation

Introduction

1. This report summarises and outlines the progress so far in producing a backseries for small area population estimates (SAPE) for Scotland by 2011 data zones for 2001-2010. Our ambition is to publish this backseries alongside the mid-2018 SAPE in August 2019.

2. The Population and Migration Statistics (PAMS) Committee is asked to note the paper, note the planned method used to adjust the estimates for consistency with published 2011 figures (as outlined in paragraphs 30-34) and offer any comments.

Small Area Population Estimates

3. The data zone is the fundamental building block for NRS population statistics and the smallest geography that NRS provides population data for outside of census years. Data zone boundaries are produced every ten years, based on populations from the Census. The most recent data zone boundaries are the 6,976 data zones created for the 2011 Census. These cover the entire area of Scotland and were designed to have populations of between 500 and 1,000 people in 2011.

4. NRS currently provides estimates of 2011 data zone populations by sex and single year of age from 2011 to 2017. As the boundaries of data zones do not change between each census, they can be used to create a comparable time series for the population of a particular area. Estimates for other areas, such as council wards and parliamentary constituencies, are approximated by adding together the populations of the data zones they contain.

5. Small area population estimates for earlier years (prior to 2011) are currently only provided for 2001 data zones, created for the 2001 census. These have similar properties to the 2011 data zones, but as they have different boundaries the population estimates for 2001 to 2010 (based on 2001 data zone boundaries) are not directly comparable with the population estimates for 2011 onwards (based on 2011 data zone boundaries).

Methodology for SAPE backseries reestimation on 2011 data zones

6. SAPE is produced using the demographic cohort component method. The data zone estimates for the previous year are aged on and adjusted for the levels of

births, deaths, net migration and other special populations (e.g. prisons, armed forces) in the area. These are then controlled to match the total population of each council area in the NRS mid-year estimates by sex and single year of age, and minor manual adjustments are made to areas with known issues such as student areas. Details on this method can be found in the methodology guide for Small Area Population Estimates[1].

7. The backseries was produced using the same methodology, starting with the 2001 Census. This required an estimate of the population of 2011 data zones at the time of the 2001 census. Generating these estimates by using best-fit matching of 2001 data zones to 2011 data zones was initially considered, but rejected due to guidance published by the Scottish Government[2]. Matching the smaller 2001 Census Output Areas to 2011 data zones was also considered, but this led to large inaccuracies due to output areas not nesting perfectly within 2011 data zone boundaries.

8. The estimates of 2011 data zone populations in 2001 were finally created by obtaining the 2001 Census population at postcode level, and then matching each 2001 postcode to a 2011 data zone. The postcode boundaries in 2001 do not nest perfectly within 2011 data zone boundaries, so this will still introduce some inaccuracies, however it was judged that these will be small enough to not have a significant impact on the population estimates.

9. Using these populations, the estimates for each data zone were rolled forward, using estimates of births, deaths, net migration, prison populations, armed forces populations, and asylum seekers by 2011 data zone for each year from 2001 to2011.

10. These rolled-forward 2011 estimates were then compared with the currently published 2011 small area population estimates based on the 2011 Census. These will be referred to as “rolled-forward” and “Census-based” estimates respectively.

11. It is assumed that the Census-based estimates are the best estimate of the population and therefore correct, and so any difference between the two is caused by an error introduced in the process of rolling forward – either an overestimate of the rolled-forward SAPE or an underestimate.

Comparison of 2011 estimates

12. As both sets of estimates were controlled to the same 2011 mid-year estimates by council area, sex and single year of age, the total population by council area, sex and age is identical in both. All differences are in the distribution of the population across data zones within each council area.

13. Over half (3,574, 51.2%) of data zones have been underestimated in the rolled-forward SAPE – with a lower population than the Census-based estimates – and

[1] https://www.nrscotland.gov.uk/files/statistics/population-estimates/sape-17/sape-17-methodology.pdf
[2] https://www2.gov.scot/Topics/Statistics/sns/SNSRef/DZMatchingQGuide

under half (3,349, 48.0%) have been overestimated. About a third of data zones (2,334, 33.5%) have a total difference of less than 20 people in either direction, and over two thirds (4,888, 70.1%) have a total difference of less than 50 people in either direction. A detailed table of differences is shown below:

Rolled-forward SAPE	Data zones	Percentage
Overestimate by 200+	61	0.9%
Overestimate by 100-199	210	3.0%
Overestimate by 50-99	692	9.9%
Overestimate by 20-49	1,263	18.1%
Overestimate by 10-19	567	8.1%
Overestimate by 1-9	556	8.0%
Exactly the same	53	0.8%
Underestimate by 1-9	565	8.1%
Underestimate by 10-19	593	8.5%
Underestimate by 20-49	1,291	18.5%
Underestimate by 50-99	842	12.1%
Underestimate by 100-199	258	3.7%
Underestimate by 200+	25	0.4%

Comparison by age and sex

14. By age, the largest adjustments are needed for young adults, particularly the 20-29 age group, as shown in Figure 1. This is primarily due to difficulties measuring the migration of students, and most data zones with large differences in these age groups are in student areas.

15. Most data zones have similar differences for both males and females. The male

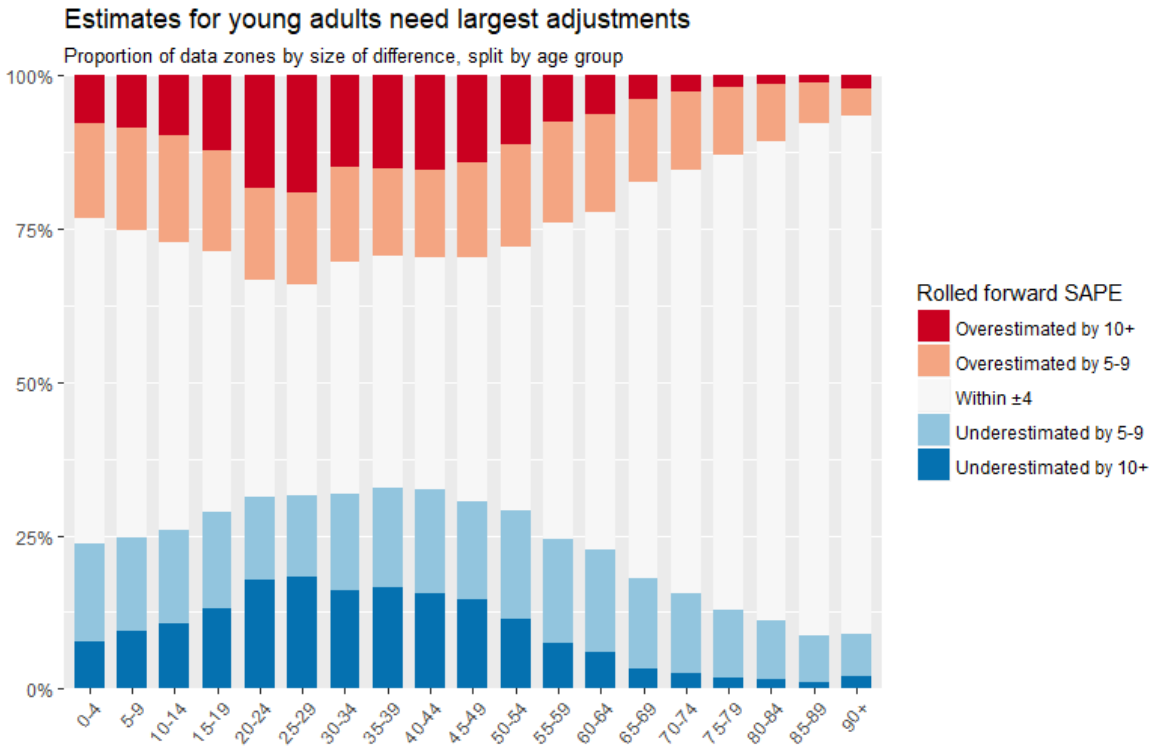

Figure 1

and female differences in 90% of data zones are within 50 people of each other, for 71% of data zones are within 30. Figure 2 shows the difference for males and females for each data zone.

16. The range for male differences is much larger than that of females, however this is mainly due to a very small number of data zones.

17. Some of the data zones shown on the graph with the largest differences are:

- S01011957 (+1,640 male, +1,080 female) – This data zone, covering areas around St Madoes in Perth, contains a farm with a large number of seasonal workers. There is a recurring problem each year where workers who migrate into the area are recorded but not all of those who leave are, leading to a very large overall overestimate. As there are more males migrating in each year, this problem applies more to males, but both sexes are affected

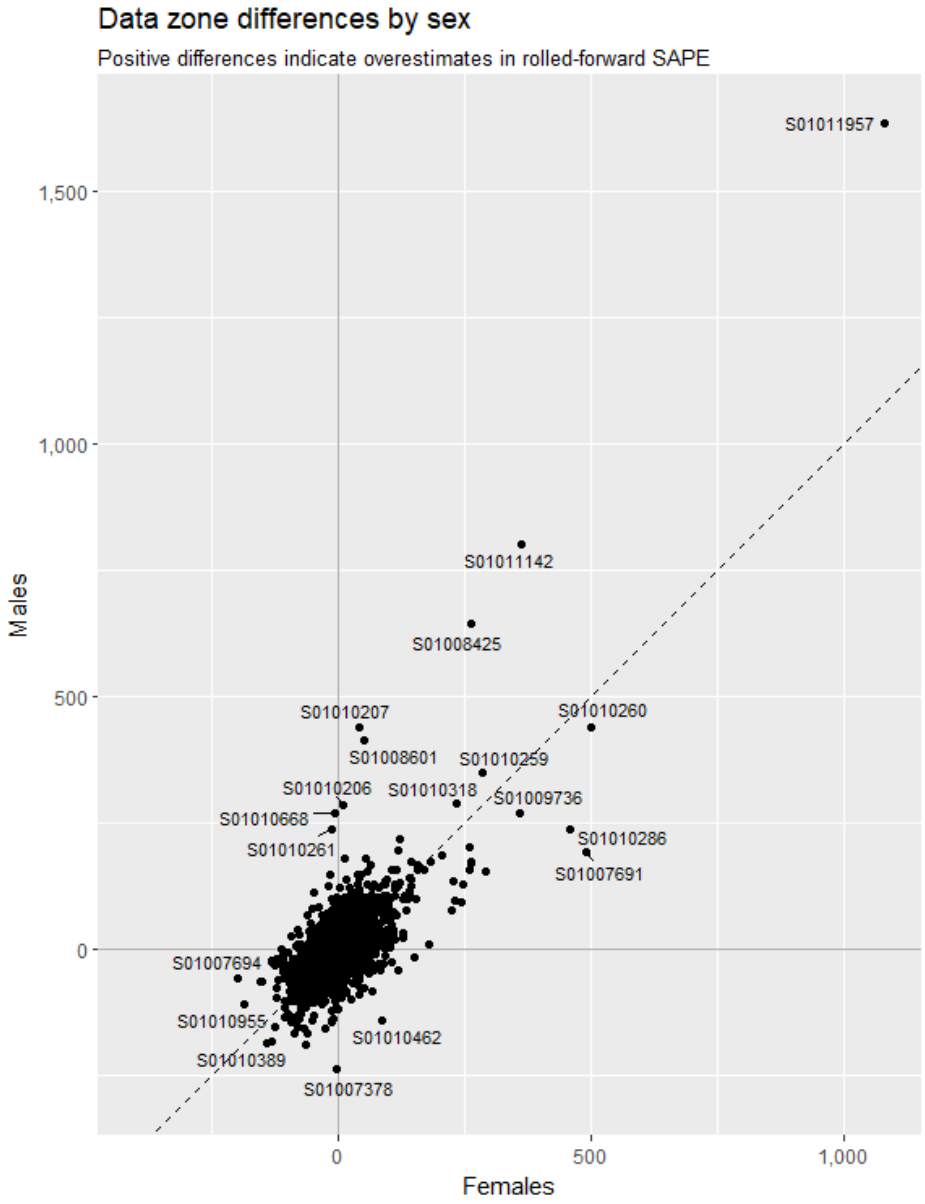

Figure 2

significantly.

- S01011142 (+800 male, +360 female) – This contains most of RAF Lossiemouth. This is due to issues with the distribution of estimates of armed forces personnel between this data zone and the neighbouring data zone S01011139, which contains the rest of RAF Lossiemouth. This affects males more due to there generally being more males in military populations.
- S01010207 (+440 male, +40 female) and S01010206 (+284 male, +9 female) – These contained the Red Road flats in Glasgow. These were being cleared of residents from 2005 in preparation for their eventual

demolition. Not all of this out-migration has been picked up, an issue which affects males more for unclear reasons.

- A number of data zones contain student halls and universities – due to the difficulties of measuring student populations large differences are expected, and student migration to small areas is often skewed between sexes. These data zones are:
  - S01008425 (+640 male, +260 female) – Riccarton campus, Heriot-Watt University
  - S01008601 (+410 male, +50 female) – Pollock Halls, University of Edinburgh
  - S01007691 (+190 male, +490 female) – City campus, University of Dundee
  - S01010261 (+235 male, -12 female) – Several halls for Glasgow Caledonian University
  - S01010286 (+240 male, +460 female) – Several halls for University of Glasgow
- S01010668 (+270 male, -10 female) – This contains the Glendoe Hydro Scheme near Fort Augustus. A large influx of male migration began in 2006 and ended in 2008, likely connected to construction of this hydroelectric facility. As with the farm near Perth, out-migration does not seem to have been properly recorded and so a large excess population remains in the 2011 estimate.
- S01007378 (-240 male, -10 female) – This contains HMNB Clyde. The population here is very large, and heavily skewed towards men (1,928 male and 377 female in the census-based 2011 estimate). While the differences are large in absolute terms, they are small in terms of the population.
- S01010462 (-140 male, +90 female) – This contains a number of tower blocks on Lincoln Avenue in Glasgow. These were being cleared for demolition, and had similar issues to the Red Road flats in S01010207.

Comparison by area

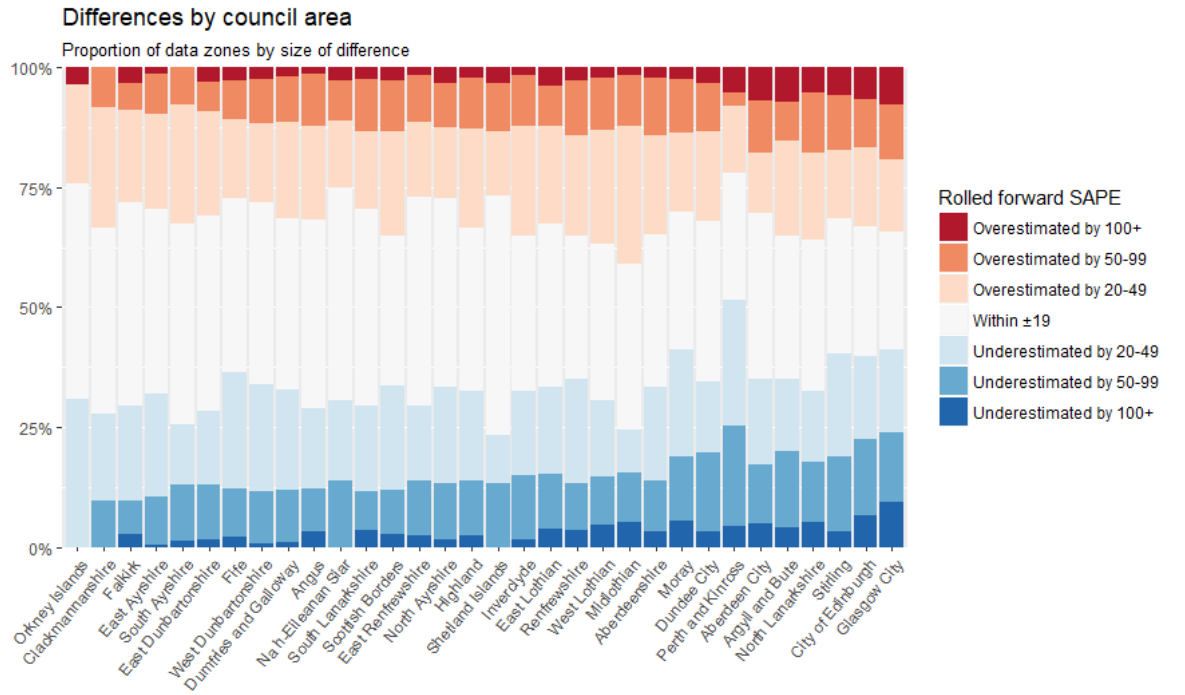

Figure 3

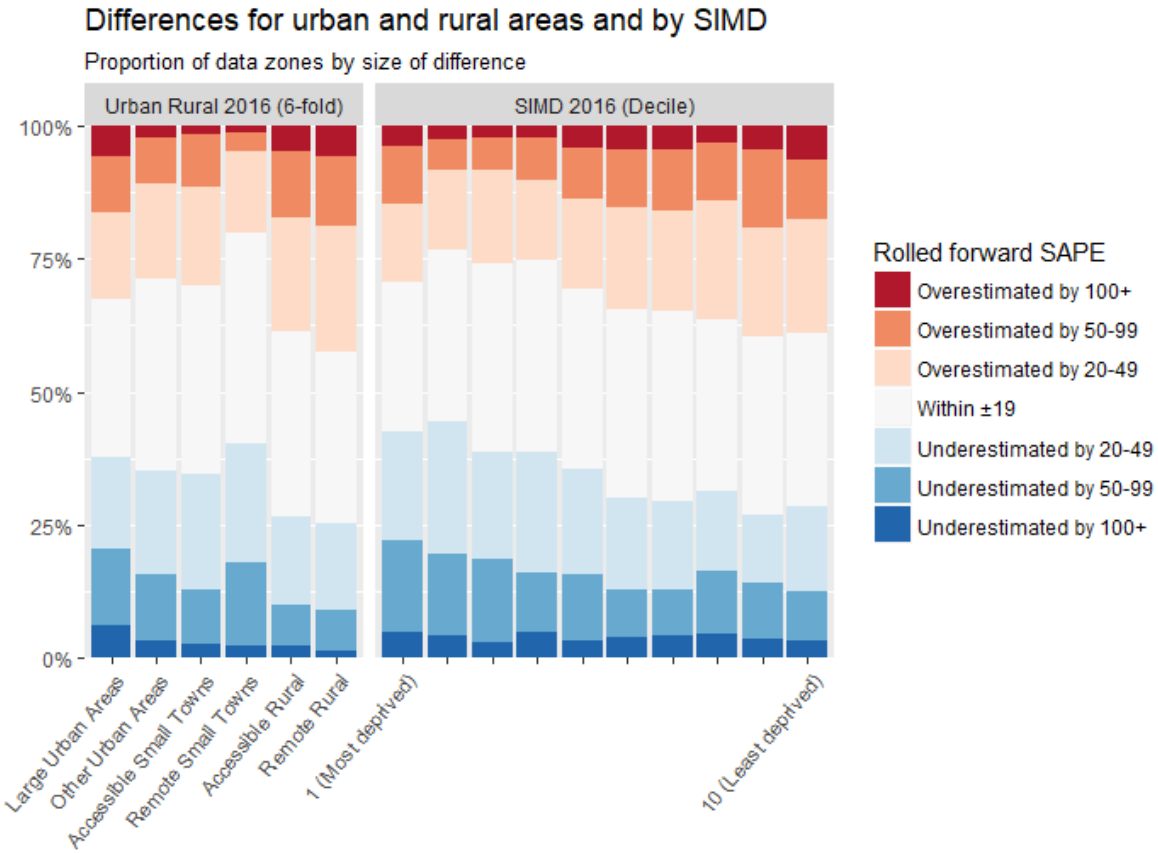

Figure 4

18. Of all council areas in Scotland, Glasgow City has the highest proportion of data zones with an overestimate of 100 people or more (58, 7.8%) and the highest proportion of data zones with an underestimate of 100 people or more (70, 9.4%). The proportions of data zones by the size of differences in 2011 is shown in Figure 3. They are shown ordered by the proportion of data zones with differences larger than 50 in either direction.

19. Figure 4 shows how the patterns of differences vary by urban and rural areas, and by deprivation of areas. The left side shows the 2016 Urban-Rural 6-fold classification (the earliest available for 2011 data zones). The Large Urban Areas category has the highest percentage of data zones with differences of 100 people or more in either direction (12.1%, 281 data zones). Rural areas also show some differences that are over 50 or over 100, but these are chiefly overestimates. Note that there are fewer rural data zones than urban data zones, and so the counts of urban areas with large overestimates are still much higher than the same counts for rural areas.

20. The right side shows patterns of differences by the Scottish Index of Multiple Deprivation (SIMD) for 2016 (the earliest available for 2011 data zones). More deprived areas generally seem to be more prone to being underestimated, while less deprived areas are more likely to be overestimated.

Proposed correction of differences

21. Correction of these differences will be carried out separately for each data zone, sex, and age cohort (those who were the same age on 30 June 2011).

Linear adjustment method

22. The initial method considered was to adjust the population linearly across time. That is, if there is an overestimate of 100 people in 2011, we will adjust down by 90 in 2010, 80 in 2009, and so on to 10 in 2002 and 0 in 2001. This method is based on that used for correcting the 2002-2010 mid-year estimates for Scotland following the 2011 Census.[3]

23. The method assumes that as we do not know where our differences have come from, they are equally likely to be from all years. So, adjustments should be equally apportioned across all years – changing each year's population in proportion to how many years of the decade have passed. An example of the results of this process in various example cohorts can be seen in Figure 5. (Note these cohorts have been chosen for their large populations and large differences in 2011 – these will be much smaller for most cohorts within data zones).

[3] For more information on this process, see the NRS publication: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/mid-2002-to-mid-2010-revision. Note that the accompanying population tables have since been superseded by corrected estimates.

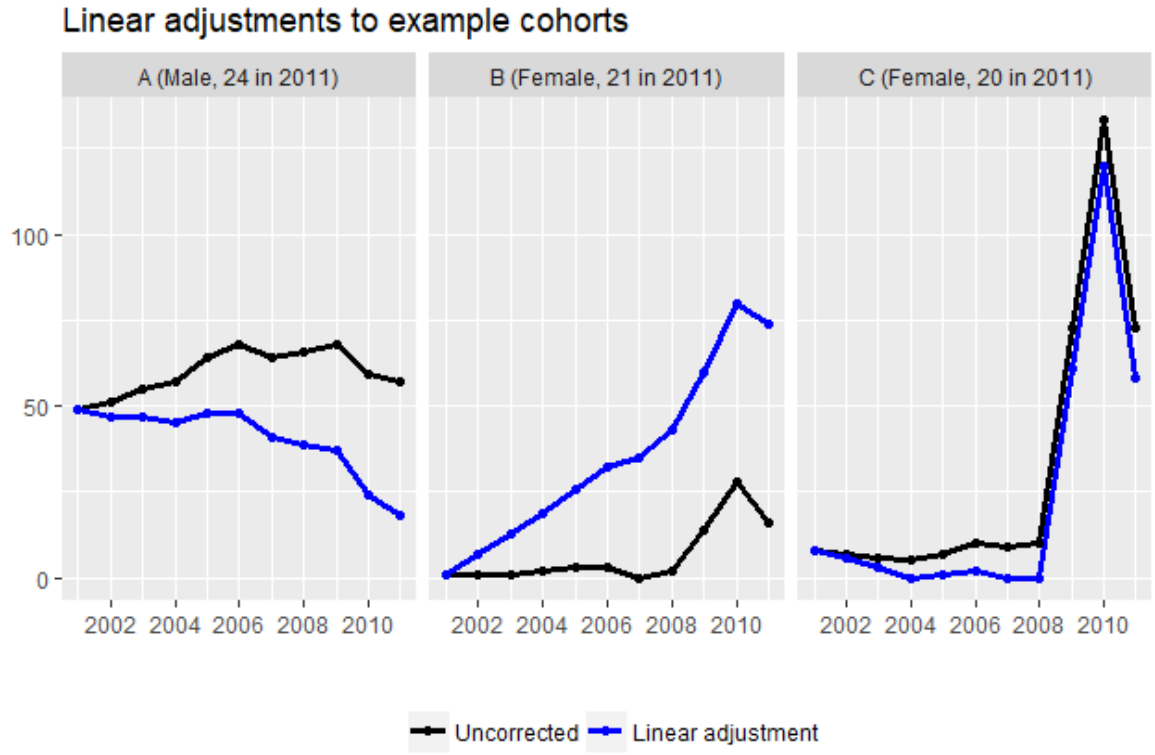

Figure 5

24. These graphs show the expected pattern of the adjusted populations starting identical to the original rebased figures, and then gradually diverging over time as the 2011 Census-based figures gain more influence, until they become equal to the Census-based figures in 2011.

25. However, closer inspection of these cohorts shows a potential problem with this method. Cohorts B and C are both in data zones near or in university campuses, where an increase in population may be expected as a cohort reaches age 17-18 (years 2007 and 2008 in both cohorts). The upward adjustments for Cohort B show a steady increase between ages 11 and 17 (2001-2007). The downward adjustments to Cohort C result in the population falling to zero between ages 14 and 17 (2004-2007). Considering the known issues with estimating student migration, It is more likely that these large differences in 2011 are due to incorrectly estimating migration of student ages, starting from approximately 17 years old.

26. This issue affects multiple data zones where there are large differences and patterns of migration suddenly changing at specific ages. Primarily these data zones are areas with a high number of students. It is normally overestimation (as in cohort B) that has the largest effect, as populations below 17 in these areas are typically close to zero to begin with.

27. Initial tests performed on data zones in the council area of Dundee City show that the controlling of all corrected estimates to the already published mid-year estimates by council, sex and single year of age do not correct this problem significantly, as the controlling method corrects for a surplus of younger ages in

a single data zone primarily by subtracting small amounts from each of a large number of other data zones with smaller populations.

Correction of issues with linear adjustment

28. One possible way of dealing with this issue is to set a cut-off date for some cohorts. For example, instead of spreading out the adjustments in cohort B equally over the ten years, they would instead be spread out between 2008 and 2011, and the estimates before 2008 kept the same.

29. This would avoid overestimation of under-17s, however it would rely on an arbitrary choice of which data zones to apply this to and when to set the cut-off at. It would also mean the adjustment procedure was inconsistent between data zones.

30. An alternative method, which we plan to use, is for each year's adjustment to depend on not only how many years of the decade have passed, but also on the total amount of change (in either direction) seen in the population so far as a proportion to the sum of all annual changes over the ten years.

31. In calculating these totals, the absolute value of the change in each year will be used so that positive and negative changes will not cancel each other out – for example, cohorts B and C undergo a dip in 2011, but this will still be counted as adding to the total change. We do not want these changes to cancel each other out as we want more adjustments to be made when the population is more volatile.

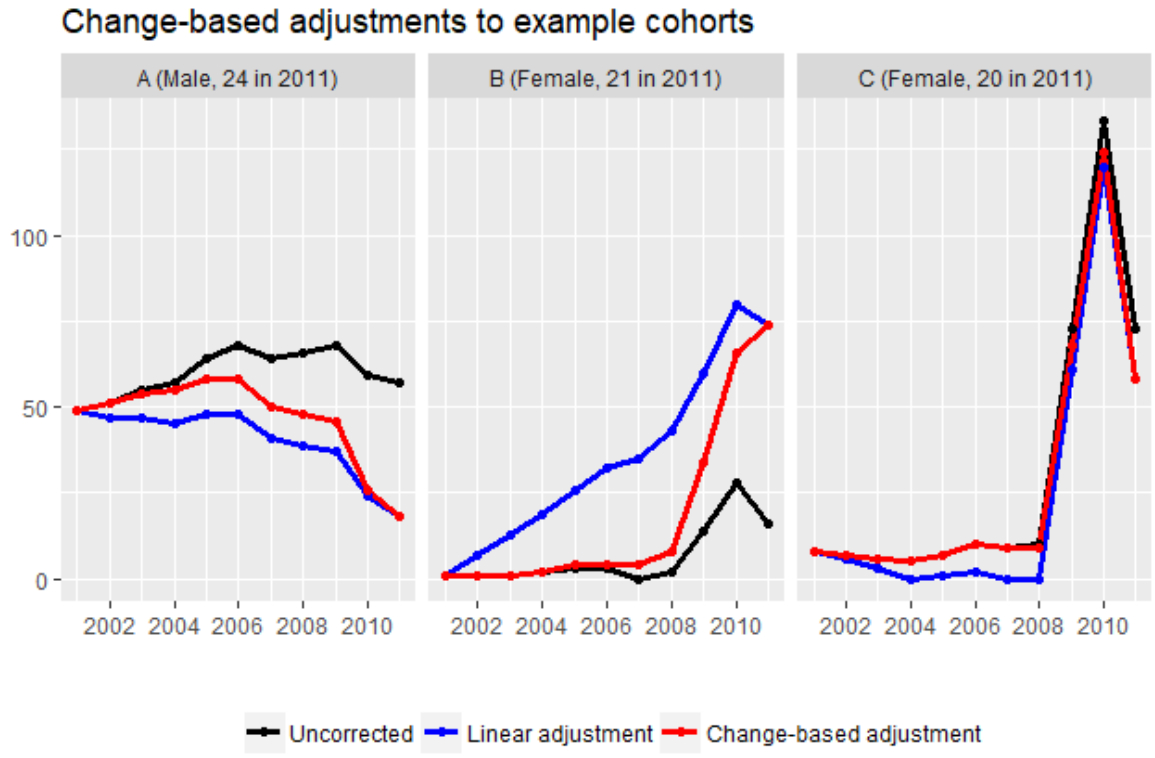

Figure 6

32. Using this method, if the rate of change in a data zone has been fairly stable, then the adjustments will be done the same as the linear method. However, if the population estimates have been more volatile from a certain year onwards, less of the adjustments will be applied to the years before.

33. An example of this method is shown in Figure 6, where it is applied to the same example cohorts as in Figure 5. In cohort B, the change-based adjustments are smaller during the earlier years, with larger adjustments taking place in 2008-2011. The lower estimates of under-17s in cohort C are avoided.

34. After applying these adjustments, the figures will be controlled to the already published mid-year estimates by council, sex and single year of age so that the totals by council area match up exactly.

NRS: Population and Migration Statistics branch
April 2019