Invitation to comment: Experimental release of the TUNDRA participation classification based on LSOAs

The accuracy of a local area classification

Overview

The Office for Students (OfS) has released higher education participation classifications POLAR4 (participation of local areas) and TUNDRA (tracking underrepresentation by area) based on the geographic areas called Mid-level Super-Output Areas (MSOAs). This document describes analysis to support the potential release of the TUNDRA participation classification based on the smaller Lower-layer Super-Output Area (LSOA) geography in England.

This document analyses the accuracy of an LSOA classification. Further information (including more context, methodology and analysis of the stability and sensitivity of the LSOA classification) is available on the OfS website.[1]

Introduction

The OfS participation classifications POLAR4 and TUNDRA suppress areas which have a small population, as they have the potential to introduce inaccuracy. In an area with a small underlying population, the chance variation of a single person changing their decision as to whether to participate in higher education or not could give rise to a large change in the participation rate and thereby an area moving between quintiles in an alarming manner.

Another way to view this is to think of the base population of students in each area as being a sample from a super-population of students in that area.

The aim of a participation classification is to understand how participation in higher education varies geographically across the country. A classification needs to accurately reflect participation from an area. TUNDRA aims for a minimum underlying population in each area (MSOA): an area which does not meet this minimum is suppressed from publication as it may not give an accurate reflection of underlying participation. In current classifications this minimum population is set at 50. This limits the potential for the chance variation of a single person changing their decision as to whether to participate in higher education or not, giving rise to the participation rate changing dramatically and the area moving between quintiles in an alarming manner. If used for an LSOA classification, suppression of data for areas with a base population of less than 50 will lead to the suppression of too many LSOAs (over 10 per cent).

[1] Available at www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-tundra/

This paper considers the effect of lowering the suppression limit to 30, which allows reporting of a greater number of LSOAs in the classification. Using simulation, the accuracy of the classification is assessed.

LSOAs which move more than one quintile are considered to show a large move, which in this context is considered an error in the classification.

A small area example

An example area is chosen which only has five students. In this area, each student is 20 per cent of the population. The participation rate can therefore be 0 per cent, 20 per cent, 40 per cent, 60 per cent, 80 per cent or 100 per cent. With zero or one participants, this area would typically fall into quintile one, whereas with three, four or five participants it would typically be quintile five. The personal choice of a very small number of students could have a huge impact on the classification. Whilst it is reasonable to believe that reporting areas this small is not appropriate, it is not clear at what underlying population we can be confident that we have assigned the correct quintile to an area. The accuracy of a classification which uses a suppression limit of 30 is investigated using simulation.

Simulation

The approach used starts from a TUNDRA LSOA classification based on students which take GCSEs in the years 2010 to 2014 inclusive. Using this as a reference, the number of participants from an area can be estimated from a binomial distribution: conceptually each person in the base population flips a (biased) coin which decides if they will participate or not. The binomial distribution needs two parameters, N and p. These are estimated from the reference classification using the LSOA base population (for N) and the LSOA participation rate (for p), for each LSOA. Every pass of the simulation therefore generates a new participation rate for every LSOA, and then a new quintile based on the new classification for all of the LSOAs.

A possible approach would be to calculate the probability of a given number of participants from the area, and use the published quintile boundaries to determine the probability of the area moving by zero, one, two, three or four quintiles. But this approach would not allow all areas to simultaneously vary at random, which also allows adjustment of the quintile boundaries.

Each classification generated by one pass of the simulation is compared to the original reference classification, to calculate the difference to the reference quintile for each LSOA. This is repeated for 2,000 simulated classifications to give stable summary statistics.

Results are reported for LSOAs grouped by the size of the base population. The percentage of LSOAs in a size band, in each quintile, is used as the denominator to calculate the percentage of those LSOAs which are expected to:

- stay in the same quintile;
- move by one quintile; or
- move by more than one quintile.

LSOAs which move more than one quintile are considered to show a large move, which in this context is considered an error in the classification.

Results

To aid interpretation of results, the number of LSOAs of each size, in each quintile, is shown. The three subsequent plots show the movement of LSOAs across quintiles, and the last plot shows the overall error rate for each size of LSOA which is then summarised in the table.

There are many ways a percentage could be defined. Starting from the number of LSOAs of a given size in a given quintile considers the question: if an LSOA of a chosen size is in a particular quintile, what is the probability of it being an error? For example, looking at an LSOA of size 30-39 in quintile one, this measure estimates the probability it being an error so that the LSOA has been allocated to a quintile which is at least two quintiles in error. This is the measure used in the graphs which follow.

LSOAs which have a population of 100 or more, are grouped into one category called ‘Population >=100’.

Figure 1: Number of LSOAs this size in the quintile

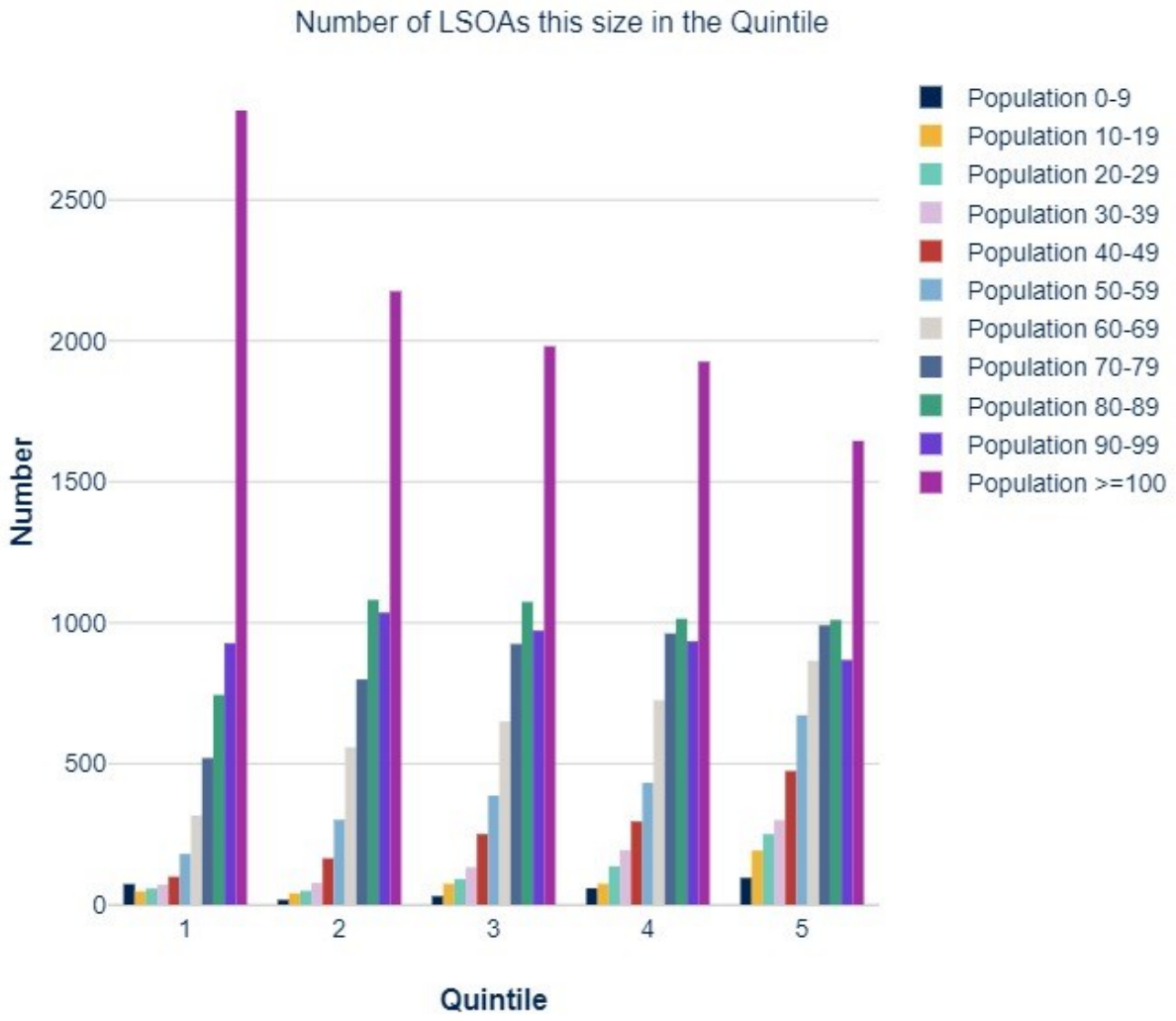

Figure 2: Percentage of LSOAs which stay in the same quintile

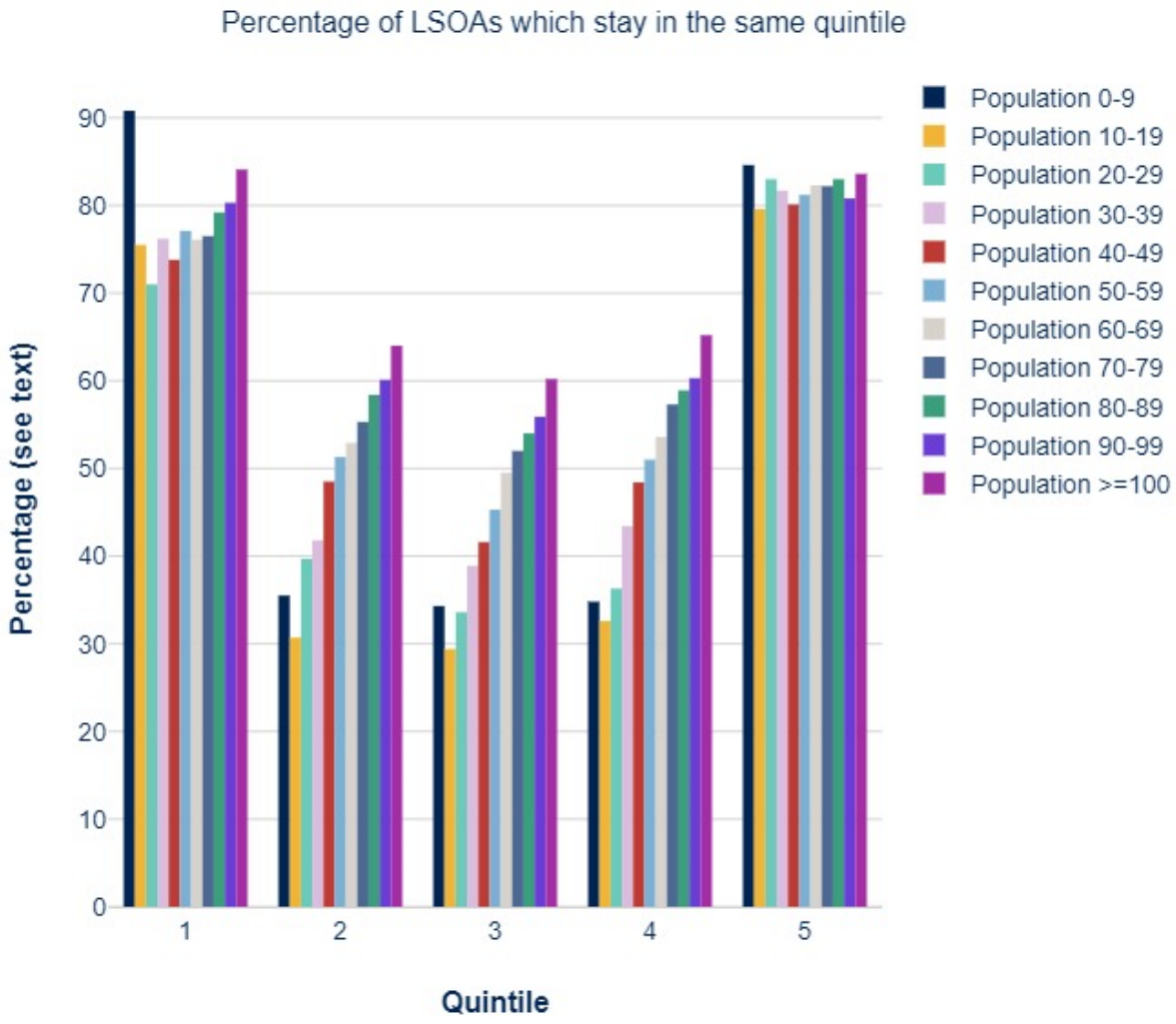

Figure 3: Percentage of LSOAs which move by one quintile

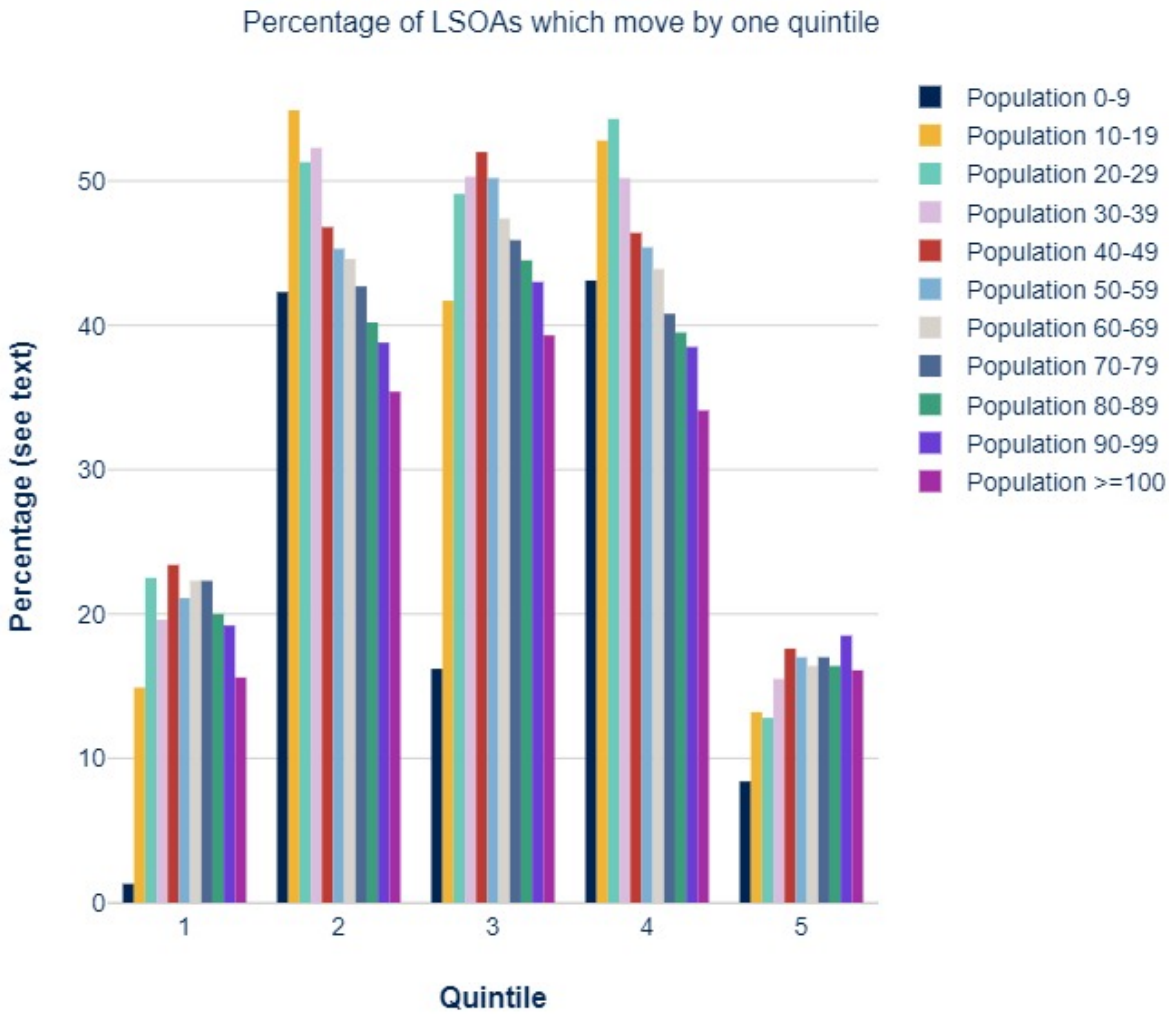

Figure 4: Percentage of LSOAs which move by more than one quintile

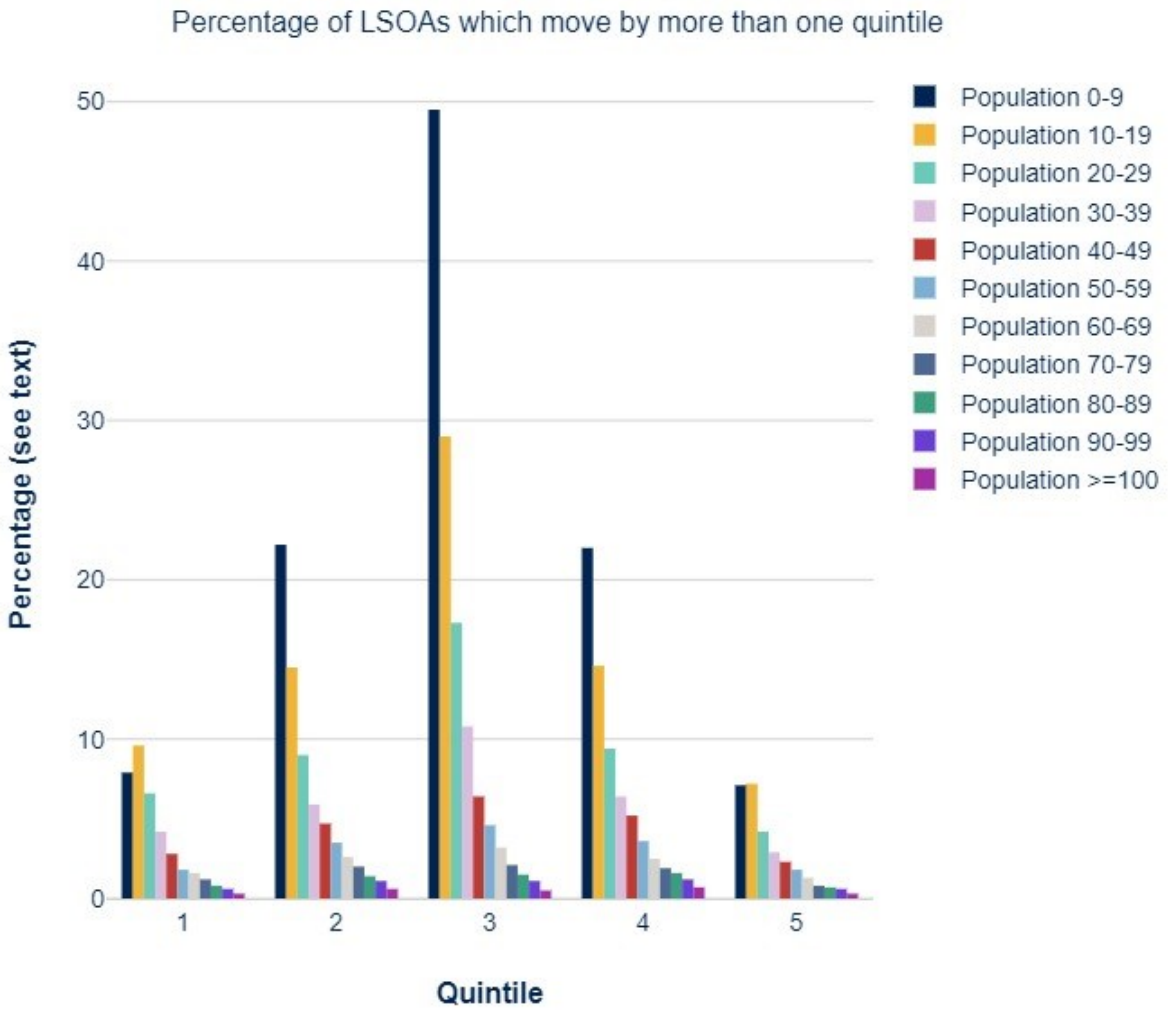

Figure 5: Percentage of LSOAs which are an error, by base population of LSOA

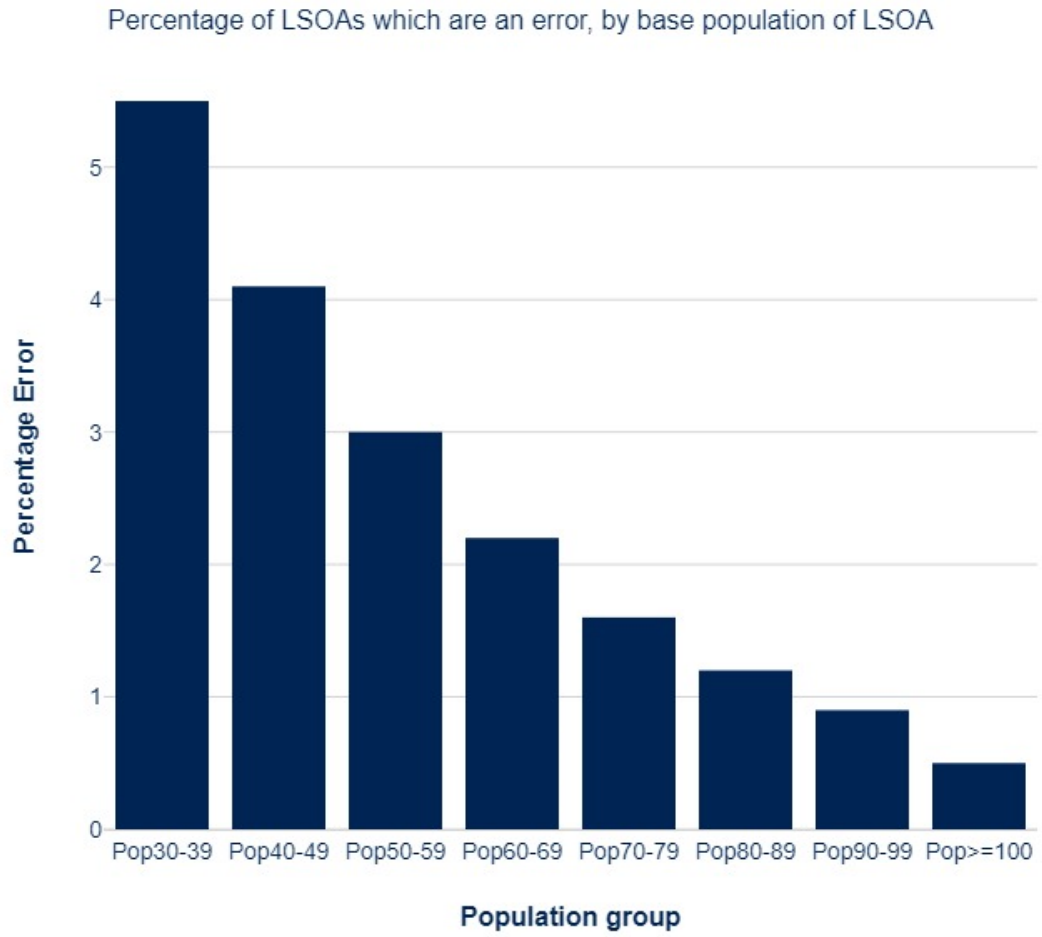

Table 1: Overall error rates, for all LSOAs of the given size in all quintiles

Population band	Percentage (%) error
30-39	5.5
40-49	4.1
50-59	3
60-69	2.2
70-79	1.6
80-89	1.2
90-99	0.9
At least 100	0.5
All LSOAs with population at least 30	1.4

For all LSOA sizes, over all quintiles, the error rate in LSOAs with a population of at least 30 is 1.4 per cent.

For the smallest population band, 30-39, the error rate overall is 5.5 per cent. This varies widely by quintile, see Table 2.

Table 2: Error rate for LSOAS with population 30-39, by quintile

Quintile	Total number of LSOAs size 30-39	Percentage error (%) for LSOAs size 30-39
1	71	4.2
2	77	5.9
3	133	10.8
4	193	6.4
5	300	2.9
Total	774	5.5

Discussion

The distribution of size of LSOAs shows the relatively small number of LSOAs with a small base population. These smaller LSOAs are more common in the higher quintiles. For LSOAs of size 30-39, in quintile five there are several times the number of LSOAs in quintile one.

The graph showing the number of LSOAs of each size, in each quintile, shows a slight dip (at size 90-99) before the last group (size at least 100). This is due to the grouping of all LSOAs with population of at least 100.

The error rate, across all quintiles, confirms the expectation that small LSOAs will be more susceptible to error than large LSOAs. The LSOAs with population 30-39 show an error rate of 5.5 per cent overall. Referring to the graph which looks at the error rate within each group, in each quintile, shows that the error rate is higher for LSOAs which are placed in quintile three than for LSOAs in any other quintile. For population 30-39, the error rate in quintile three is over twice the error rate in quintile one; and the error rate in quintile one is, in turn, much larger than the error rate in quintile five. In this way an overall error rate hides important information which is tied up in the relationship of error rate to quintile, and the distribution of the size of LSOAs across quintiles.

The number of LSOAs of size 30-39 in each quintile, and the error rate for those LSOAs, is shown in Table 2. In quintile one, there are 71 LSOAs size 30-39, with an error rate of 4.2 per cent. This suggests an expectation that three of these will be reported in error.

The overall error rate for all LSOAs in the classification is 1.4 per cent. There are 32,844 LSOAs in England (note that not all of these will be reported due to low population), and 444 LSOAs are expected to be reported in error overall.

The patterns depicting movement of LSOAs by size and quintile are generally as would be expected:

- smaller LSOAs tend to move quintile more than larger LSOAs
- quintiles one and five are much wider (cover a larger range of possible participation rates) than quintiles two to four, so tend to show less movement than quintiles two to four

- quintile five is much wider than quintile one, so tends to show less movement
- quintile three shows a larger proportion of large moves compared to other quintiles. An LSOA can move both up two quintiles, or down two quintiles from quintile three. An LSOA in quintile two could move up by two quintiles, but only ever move down by one quintile as there is nothing lower than quintile one
- the very small LSOAs show some unexpected behaviour, note there are relatively few in the dataset and the granularity of a one-person random variation can be very large.

Conclusion

Reporting results for LSOAs with population of at least 30 limits the overall error rate to 1.4 per cent. However this hides important information. The error rate is higher for the smaller LSOAs: the group of size 30-39 show an overall error rate of 5.5 per cent. But within this group, both the distribution of LSOAs across quintiles, and the error rate within the quintile, vary widely. There are 71 LSOAs this size in quintile one in the classification, with an expectation that about three of these (4.2 per cent) will be reported in error.