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

The stability and sensitivity of a local area classification: Overview

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

For information about our area-based measures, see the OfS website.[1]

This document analyses the stability of an LSOA classification. Further information (including more context, methodology and analysis of the accuracy of the LSOA classification) is available on the OfS website.[2]

The aim of this document is to present the results of our analysis to help inform feedback. Only when implications are considered clearly unacceptable, is an opinion included in the text.

Introduction

The aim of a participation classification is to understand how participation in higher education varies geographically across England.

A participation classification should help to identify areas where there is persistent low (or high) participation in higher education. As a classification is based on past data, there is the assumption that the future participation in an area will be well-predicted by past participation: this is known as stability. If there is no underlying change in the participation rate, a stable classification will give a good predication of future participation.

Conversely it is desirable to be able to see the effect of any change in the participation rate. For example when understanding the success of an outreach program it may be highly desirable to design a classification which is sensitive to changes in the participation rate and will therefore indicate the success of the program.

An ideal classification would be stable if there is no change in the underlying participation rate, whilst being sensitive to an actual change in the underlying participation rate. In practice there will

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

be a trade-off between these: increasing the sensitivity of a classification makes it susceptible to random variation in the actual number of participants which makes the classification less stable.

Both POLAR4 and the current TUNDRA release are based on the MSOA geography, requires a minimum population of 50 for an area to be reported, and combine five cohorts of students. For a classification based on the LSOA geography, a suppression limit of 50 would lead to a large number of LSOAs (about 10 per cent of LSOAs) being suppressed in publication. Reducing the suppression limit will include more areas, but increases the error rate (the random variation of quintile for an LSOA, see our accuracy document).[3] A suppression limit of 30 has an overall error rate to close to five per cent, and allows publication of a much larger percentage of LSOAs than a suppression limit of 50 as it suppresses less than four per cent of LSOAs. Analysis of an LSOA classification therefore initially considers suppression at 50 in order to compare to the MSOA classification with suppression at 50, and then reducing this suppression limit for the LSOA classification to 30.

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

Moving from MSOAs to LSOAs will give a smaller population for each area: this is one aspect of the experimental LSOA classification. In order to understand the effect of moving to the smaller LSOAs, analysis of the LSOA classification first follows the MSOA methodology, requires a minimum population of 50 for an area to be reported, and combines five cohorts of students.

The second aspect of moving from the MSOA to LSOA classification is, moving from suppression below 50 to suppression below 30. This is investigated by comparing an LSOA classification with suppression below 50, to an LSOA classification with suppression below 30. This gives an indication of the increased number of LSOAs which change quintile over time.

Any changes between two classifications, over time, could be due to an underlying change in the participation rate or due to instability due to the smaller suppression limit. Simulation allows separation of these two effects, enabling us to ascertain the proportion of LSOAs which move due to a change in underlying participation, and the proportion which move due to instability.

There will be a delay in calculating and publishing a TUNDRA classification for the latest available data, regardless of whether it is based on MSOAs or LSOAs. Currently (in 2020) the OfS receives the relevant higher education data up to 18 months after the start of the higher education academic year. If the TUNDRA classification were to be published as soon as this data is available to the OfS, it would already be 18 months out of date compared to a classification based on data at that point in time if it were to be available. The earliest academic year when a classification could influence targeting of outreach is about two academic years after data for the last group starting higher education in the available classification. Analysis therefore emphasises the two-year timeframe.

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

Analysis

The TUNDRA MSOA classification

The time-series of TUNDRA classifications are labelled according to the (GCSE) National Pupil Database (NPD) years which form the base population for an area. Here, the GCSE summer is in the later part of the academic year, so that students from the academic year 2006-07 take GCSEs in 2007.

Table 1: Data included in each TUNDA LSOA classification

TUNDRA version	NPD GCSE summer (inclusive)	Years for including participation in higher education at age 18 or 19 (inclusive)
0711	2007 to 2011	2009 to 2014
0812	2008 to 2012	2010 to 2015
0913	2009 to 2013	2011 to 2016
1014	2010 to 2014	2012 to 2017

A time-series of four classifications is created. The OfS receives data from both the Higher Education Statistics Agency (HESA), and the Individual Learner Record (ILR) which gives information concerning higher education in the further education sector. Since 2009 there has been a single OfS methodology for processing data from both HESA (the HESA student, and HESA Alternative records), and the ILR data, which makes this a natural choice for the first year of higher education entry. The first year of GCSE data (2007) counts back from this.

Starting from the 0711 classification, it is possible to look one year later and identify which MSOAs move in the 0812 classification and how far they move. Again starting from the 0711 classification it is possible to look two years later (to the 0913 classification) and again identify how many MSOAs move and how many quintiles they move. Last, comparing the 0711 and the 1014 classifications generates a picture of how many MSOAs move and how far, over a three-year time-lag.

The four MSOA-based classifications are generated, with suppression below a population of 50 reflecting the current methodology for publishing the classification. As there will be some random fluctuation in the definition of quintiles, in line with previous OfS and Higher Education Funding Council for England (HEFCE) analysis an MSOA needs to move at least two quintiles to be considered a ‘large’ move.

Results are given in Table 2.

Table 2: Summary of stability of TUNDRA MSOA classification

Time difference (TUNDRA versions)	Number MSOAs with large move	Percentage (%) of total number of MSOAs
One year (0711 series to 0812 series)	0	0
Two years (0711 series to 0913 series)	1	0.01
Three years (0711 series to 1014 series)	21	0.31

Imagine planning outreach in late 2016. At this time, due to the delay in data availability, the most recent classification available would be the one based on the 0711 NPD data. When planning, you would like to know if the 0711 classification still reflects young participation in 2016. With the benefit of hindsight it is now possible to look at the comparison of the TUNDRA 0711 classification to the TUNDRA 0913 classification, which is the first classification to include the 2016 entries to higher education. We conclude that only one MSOA in England moved more than one quintile. This suggests that planning based on the 0711 classification would be reliable in late 2016.

Now imagine planning a two-year programme starting in late 2016, which will last for both 2016 and 2017. The 0711 classification gives a good prediction of the quintile which will be assigned to an MSOA in 2016. Comparing the 0711 classification to the 1014 classification (which would include the latest data, if it were available at that time), shows that the 0711 quintile is a good prediction of the 1014 quintile in all but 21 MSOAs (0.31 per cent): this is about one in 300 MSOAs where the 0711 classification differs substantially to the most up-to-date classification. Again, planning based on the TUNDRA 0711 classification would be reliable for 2017.

Using an MSOA classification one, two or three years after the latest data will give a reliable picture of participation at the later date. Conversely, within the three-year time frame there will be areas where there is a genuine underlying change in the participation rate. Given large investment in outreach, some of which is closely targeted, it would be encouraging to see this reflected in the classification. The low number of areas which move more than one quintile over three years, suggests this is unlikely.

Stability and sensitivity of the TUNDRA LSOA classification

To understand the changes introduced by moving to a classification based on LSOAs, investigation first uses a similar methodology to the MSOA classification with suppression of an area with a population below 50 and combining five cohorts of students. This will give an indication of the amount of variation which is introduced by moving to the smaller areas, with all other characteristics of TUNDRA held constant.

A second investigation considers the effect of reducing the suppression cut-off to require a (five-year) base population of only 30 in an LSOA. Comparing this to the LSOA classification suppressing below 50 shows how much stability is lost when reducing the suppression limit: additional changes in this classification suggest the proportion of areas which are expected to move due to the smaller suppression population.

Note that the LSOAs which have a population below the suppression limit are included in the classification methodology and are suppressed at publication. The two underlying classifications with suppression at 50 and suppression at 30, are exactly the same but would be reported differently (changing the suppression limit before classification of the LSOAs could alter the quintile allocated to some LSOAs).

LSOA classification using TUNDRA MSOA methodology

The TUNDRA classification is generated for LSOAs based on suppression below a population of 50, mirroring the methodology used for the published version of TUNDRA based on MSOAs. The number of LSOAs in each (population-weighted) quintile, before suppression, is given in the table. There are six LSOAs which have no eligible base population over the whole of the five cohorts.

Table 3: Number of LSOAs in each quintile of the TUNDRA LSOA 0711 classification before suppression

Quintile	Number of LSOAs
1	5,851
2	6,298
3	6,567
4	6,749
5	7,362
Undefined (zero population)	6

Stability is assessed similarly to the MSOA classification: by comparing the 0711 classification to the 0812 classification (a one-year delay), the 0913 classification (a two-year delay), and the 1014 classification (a three-year delay). The number of LSOAs which move by more than one quintile is reported.

Table 4: Summary of stability of TUNDRA LSOA classification with suppression below 50

Time difference (TUNDRA LSOA versions)	Number of LSOAs which move more than one quintile	Percentage (%) of LSOAs which move by more than one quintile, out of total number of LSOAs
One year (0711 to 0812)	39	0.1
Two years (0711 to 0913)	343	1.0
Three years (0711 to 1014)	889	2.7

This LSOA classification is presented here to compare stability to the similar MSOA classification, and gives an indication of changes in the classification which may be attributed to use of the smaller LSOAs. The LSOA classification with suppression below a population of 50 is not

considered suitable for publication, due to the large proportion (about 10 per cent) of LSOAs which will be suppressed by requiring at least 50 students in an area.

LSOA classification using suppression at base population of 30

The second LSOA classification considered suppresses areas with a base population below 30. This suppresses 1,257 LSOAs (3.9 per cent) in the 1014 classification which is felt to be more acceptable in practice than the larger number of LSOAs which would not be reported when suppressing below 50.

The 0711 classification is compared to the 0812 classification (a one-year delay), the 0913 classification (a two-year delay), and the 1014 classification (a three-year delay). The number of LSOAs which move by more than one quintile is given in Table 5.

Table 5: Summary of stability of TUNDRA LSOA classification with suppression below 30

Time difference (TUNDRA versions)	Number of LSOAs which move by more than one quintile	Percentage (%) of LSOAs which move by more than one quintile, out of total number of LSOAs
One year (0711 to 0812)	55	0.2
Two years (0711 to 0913)	439	1.3
Three years (0711 to 1014)	1,101	3.4

Changes in classification and the smaller suppression limit

In England, there are just under five LSOAs in each MSOA. The MSOA series has one MSOA with a large move after two years which broadly suggests about five LSOAs would be expected to show a large move over the same time delay. However the LSOA series with suppression at 50 has 343 LSOAs which move a long way over two years. A similar comparison for a three-year delay confirms the increased movement shown by the LSOA classification with a suppression limit of 50, when compared to the MSOA classification with a suppression limit of 50.

The series with 30 as a suppression limit necessarily contains all the LSOAs with 50 as a suppression limit, as suppression takes place after the definition of quintiles. Out of the 439 LSOAs with a large move two years later and population of at least 30, 343 will have a population of at least 50. The remaining 96 LSOAs with a large move over two years have population between 30 and 50.

The rough estimates in the rest of this section use the ratio of about five LSOAs in each MSOA. Suppose we randomly choose 60 MSOAs – this is just under one percent of MSOAs, and is a convenient size to imagine. The size of 60 has no practical significance.

For the two-year delay between series, very rough estimates suggest:

- in every 60 MSOAs you would expect to find about three LSOAs showing a large move whilst having a population of at least 50 (the 60 MSOAs are about 300 LSOAs)

- within the same 60 MSOAs, you would expect to find one more LSOA which show a large move. This will have a base population between 30 and 50.

Comparing the movement in classifications with a three-year delay suggest that moving to an LSOA-based classification will again account for the majority of large moves: only about one-fifth of large moves (202 out of 1,101) are from LSOAs with population smaller than 50. The rest of the large moves (889 out of 1,101) being from LSOAs with a population of at least 50.

About 3.4 per cent (1,101) of LSOAs show a large move over three years with suppression at 30. If these are spread across MSOAs, then similar rough estimates to those above suggest about one in six MSOAs will contain an LSOA which shows a large move over three years: about four-fifths of the LSOAs which show a large move over three years will have a population of at least 50.

In summary for the three-year delay between series:

- in every 60 MSOAs you would expect to find about 10 LSOAs showing a large move whilst having a population of at least 50

- within the same 60 MSOAs, you would expect to find two further LSOAs which show a large move. They will have a base population between 30 and 50.

How much of the movement is random?

As noted in the introduction, there is a balance between stability and sensitivity. An ideal classification methodology would only allow an LSOA to move to another quintile if the underlying participation rate has changed. In practice, there will be some errors around the quintile of an LSOA due to random noise. From the estimates above, it is not possible to say what proportion of the large moves are expected due to random noise. Conversely it is not possible to say what proportion of the large moves are likely due to a change in the underlying participation rate.

The paper about accuracy[4] considers the variation in quintile expected due to random variation of participation given a known underlying participation rate and base population for each area. The simulation techniques used are now adapted to suggest the proportion of movement which is expected due to chance variation, when moving from the 0711 classification to the later ones with a one, two, or three-year delay.

Description of simulation

Starting from the 0711 TUNDRA and moving to the 0812 classification, within each LSOA the 2007 GCSE cohort will drop from the classification and the 2012 GCSE cohort will enter the classification. In this way only around one-fifth of the underlying dataset will change from the 0711 classification to 0812. Four of the five cohorts of students remain in the classification, in both the base population and the number of participants for that LSOA.

This change of part of the dataset can be modelled using simulation, based on a similar method to the accuracy paper. 0711 is considered the baseline dataset and classification with all differences shown in relation to 0711. Moving to the 0812 classification, in each LSOA the base population and number of participants from 2007 is removed from the data. The number of students in the LSOA for 2012 are included in the new base population for the LSOA. The extra number of participants

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

for the LSOA is generated from a binomial distribution based on the number of extra students in the 2012 base population, and the participation rate from the 0711 classification. The classification is generated for the new dataset, and differences from the 0711 classification are summarised. This is repeated for 2,000 simulations, following the methodology for the accuracy evaluation with suppression below 30. The mean number of LSOAs which move more than one quintile is calculated: with 2,000 replications this is stable to at least one decimal place for a base population of at least 30. For this analysis which reports LSOAs of all sizes (including less than 30), the estimates may not be this stable for populations below 30. This will counteract the effect of increased stability when only a part of the population and number of participants is changing (in the accuracy evaluation, the number of participants for the whole base population is varied). The results are given for all population sizes, illustrating the pattern in changes for large and small population sizes, bearing this in mind.

The simulation gives a picture of how much random variation would be expected in the classification when including one extra year of students, if there is no change in the underlying participation rate.

The changes suggested by the simulation can be compared directly to the changes seen when moving from the 0711 classification to the 0812 classification. The comparison gives a strong indication of how much variation is due to random noise, implying how much variation may be attributed to a change in the underlying participation rate.

Simulation results for the one-year delay

From each base quintile, the number of LSOAs in each size group is calculated. The percentage of these which show no move, a one quintile move, and a large move are respectively shown on graphs.

When interpreting the graphs, note that quintiles one and five are a lot wider (have a larger range of participation rates) than quintiles two, three and four, since quintile one starts at 0 per cent participation and quintile five ends at 100 per cent participation.

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 showing this particular type of move? This is the measure used in the graphs which follow.

Figure 1: Percentage of LSOAs which have no move one year after data

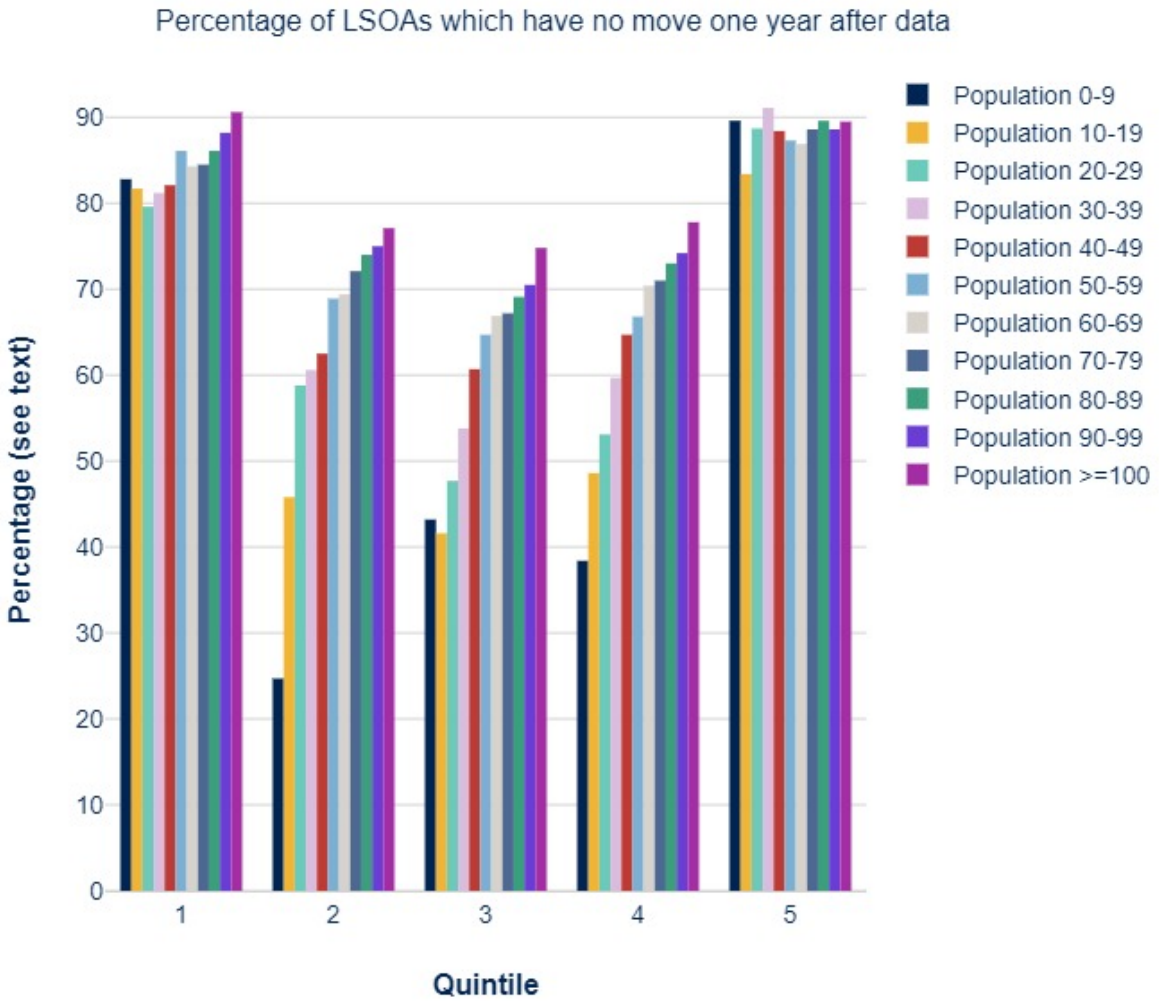


Figure 2: Percentage of LSOAs which move by one quintile one year after data

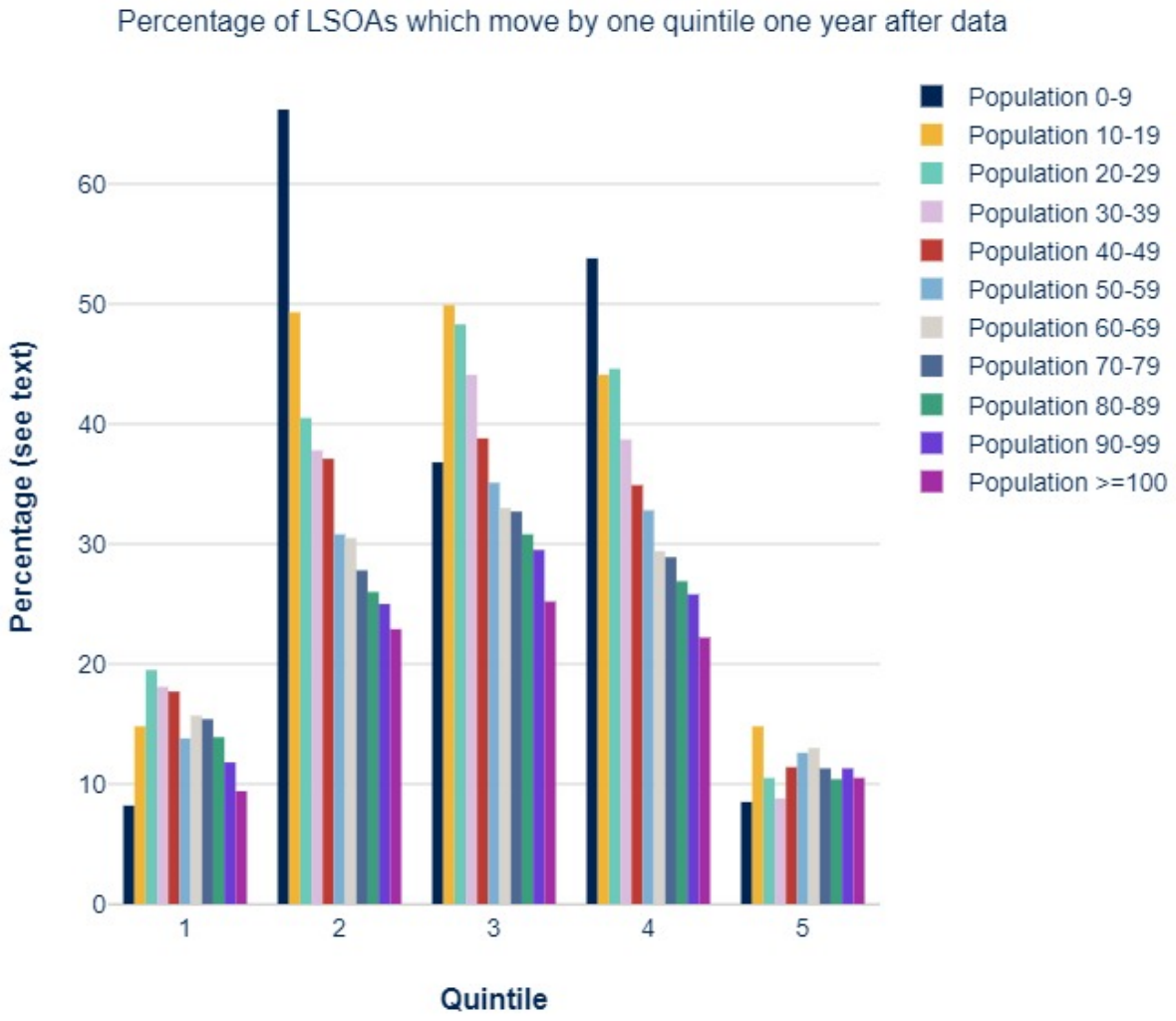


Figure 3: Percentage of LSOAs which move by more than one quintile one year after data

The simulations for a one-year delay suggest patterns which are later continued for the simulations for a two and three-year delay.

- Increasing the base population in an LSOA means it is more likely to stay in the original quintile and less likely to show a large move.
- There tends to be more movement from quintiles two, three and four, than from quintiles one or five. This is believed to be due to quintiles one and five being a lot wider than quintiles two, three and four, as quintile one starts at 0 per cent participation and quintile five ends at 100 per cent participation.
- The small population size groups show some apparently odd patterns. This is believed due to the relative granularity of the participation rate for smaller population sizes: for a base population of 10, each participant is a change of 10 per cent to the participation rate. This can result in a one-person participant change giving a greater-than-one quintile change for the LSOA, with a single-quintile move for a particular LSOA maybe not being possible.
- For quintile five the single quintile moves do not show any apparent pattern – this changes for the larger delays and shows a similar effect to quintile one noted above.

Simulation of the two-year time delay

A similar simulation is run for a two-year delay. Here, the 2007 and 2008 populations are removed from the base population, and the 2012 and 2013 populations are now included. The number of participants for the two new years in the dataset (2012 and 2013) is a sampled as a binomial value using the known base population for the two years, and the participation rate for the 0711 classification. The total number of participants is calculated by adding this to the known sum of participants for 2009, 2010 and 2011. This has introduced a greater random part to the number of participants in comparison to the one-year delay.

Figure 4: Percentage of LSOAs which have no move two years after data

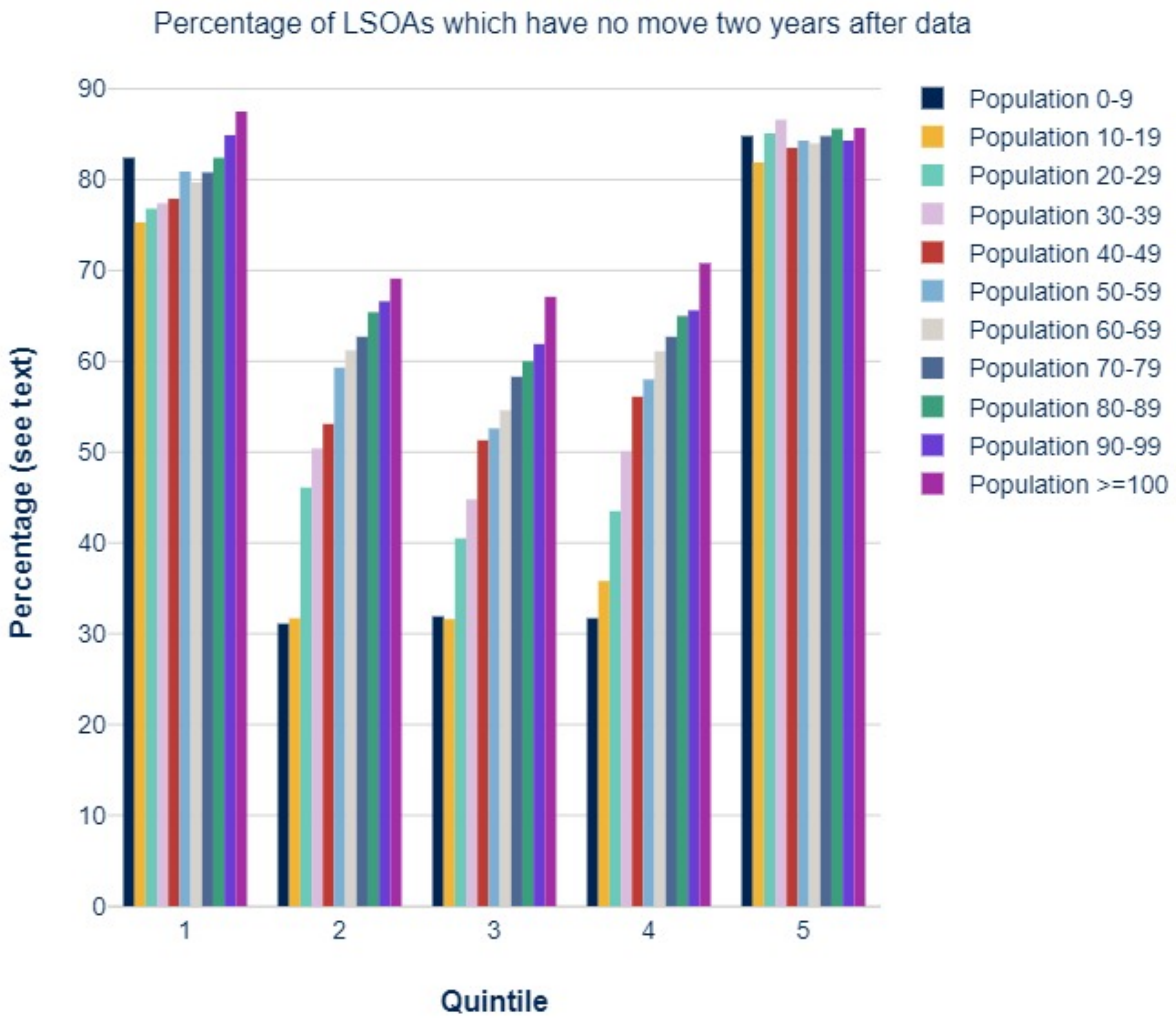


Figure 5: Percentage of LSOAs which move by one quintile two years after data

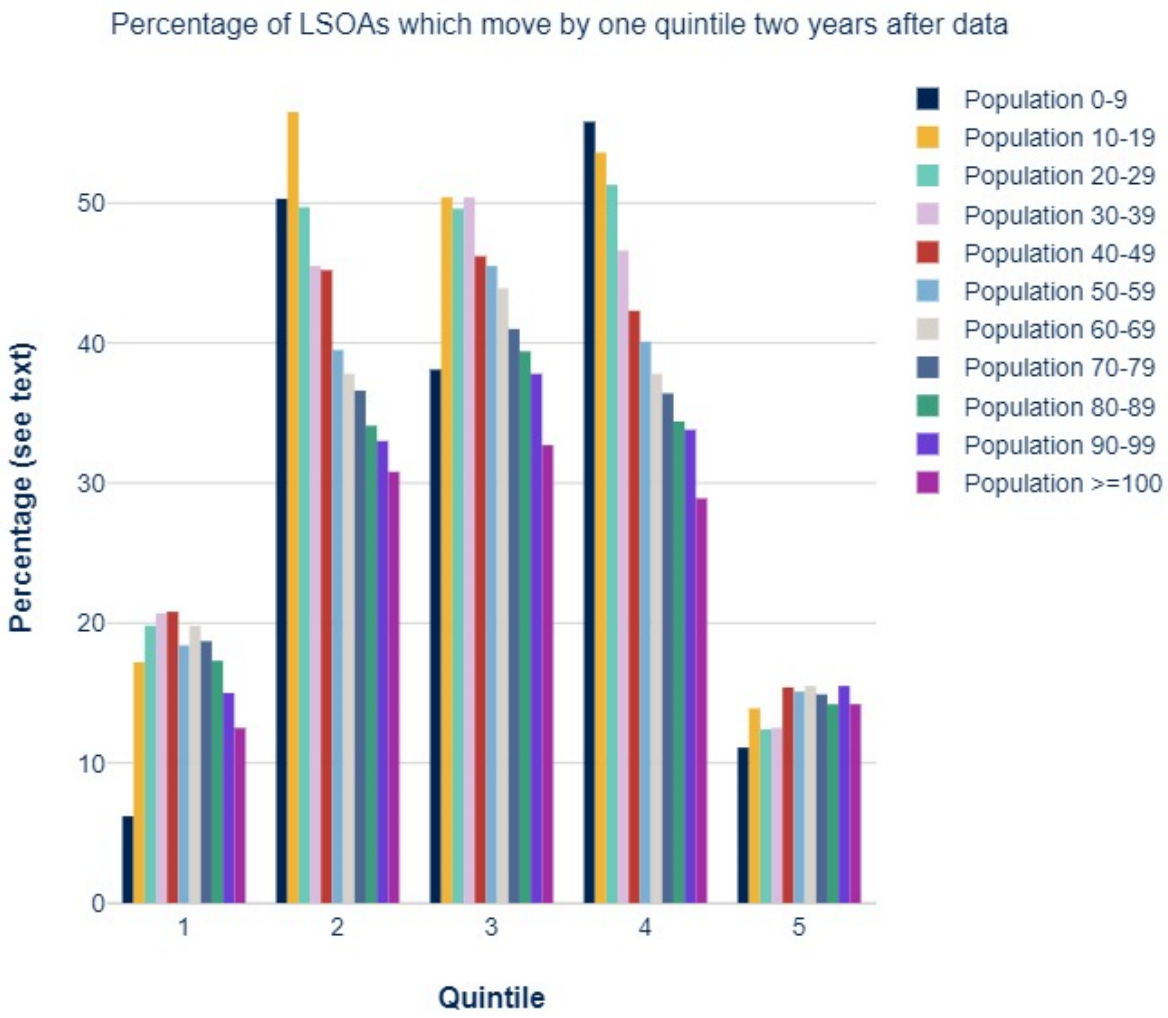
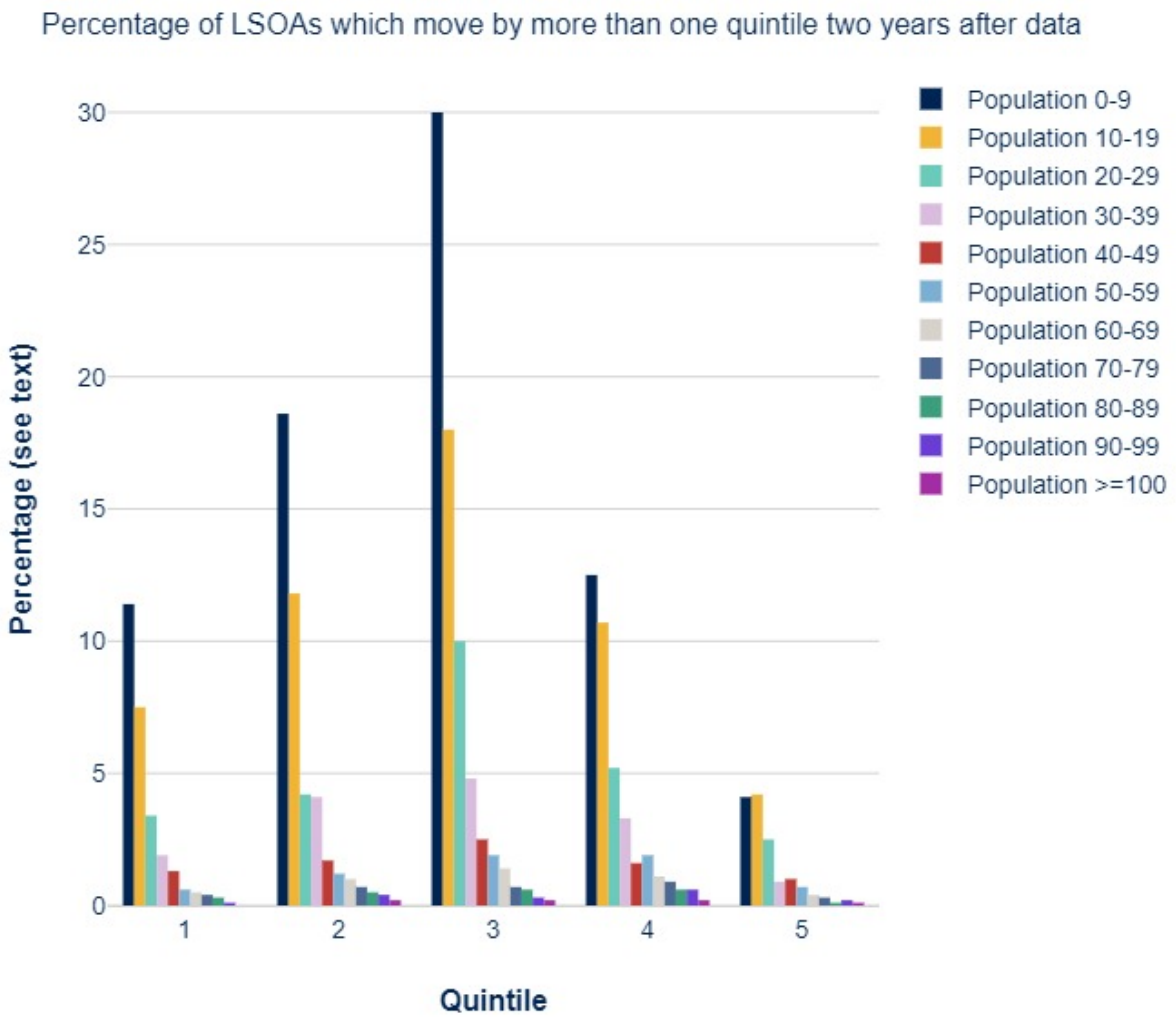


Figure 6: Percentage of LSOAs which move by more than one quintile two years after data



The two-year delay simulation results can be compared to the variation seen in practice. Recall that moving from the 0711 classification to the 0913 classification results in 439 (1.3 per cent of total) LSOAs showing a large move, of which 343 have population of at least 50. The simulation suggests that due to random variation, on average 169 (0.54 per cent of total) LSOAs would show a large move, comprised of 129 (0.41 per cent of total) LSOAs of size at least 50, and a further 40 (0.13 per cent of total) LSOAs size 30-49. The difference between these values suggests there is a number of LSOAs which show a large move and this would not be expected due to random variation. These ‘excess’ large moves are attributed to sensitivity and an underlying change to the participation rate in these LSOAs. The information is summarised in Table 6.

Table 6: Summary of stability and sensitivity results for a two-year delay

Size of LSOA	Population 30-49	Population at least 50	Population at least 30
Number of large moves in practice	96	343	439
Number of large moves expected due to random variation from simulation	40	129	169
Percentage (%) of large moves in practice which are expected due to random variation	41.7	37.6	38.5
Number of large moves expected due to sensitivity	56	214	270
Percentage (%) of large moves in practice which are attributed to sensitivity	58.3	62.4	61.5

Extending the rough estimates earlier, again choose 60 MSOAs at random. This is just under one per cent of MSOAs in England and a convenient number to imagine – the number 60 has no practical significance here.

For the two-year delay between series (note rounding errors here):

- in every 60 MSOAs you would expect to find three LSOAs showing a large move whilst having a population of at least 50. Of these three LSOAs, one is expected to be due to random variation and the other two are expected to be due to sensitivity
- in every 60 MSOAs you would expect to find one further LSOA which shows a large move. This will have a base population between 30 and 50. Just under half the time this is expected to be due to random variation, the remaining time it is expected to be due to sensitivity.

Simulation of the three-year time delay

A further simulation considers a three-year delay. Here, the 2007, 2008 and 2009 populations are removed from the base population for each LSOA, and the 2012, 2013 and 2014 populations are now included. The number of participants for the three new years (2012, 2013 and 2014) is a sampled as a binomial random variable using the known base population for the three new years, and the participation rate for the 0711 classification. The total number of participants is calculated by adding this to the known sum of participants for 2010 and 2011. This has introduced a further random part to the number of participants in comparison to both the one-year delay and the two-year delay.

Figure 7: Percentage of LSOAs which have no move three years after data

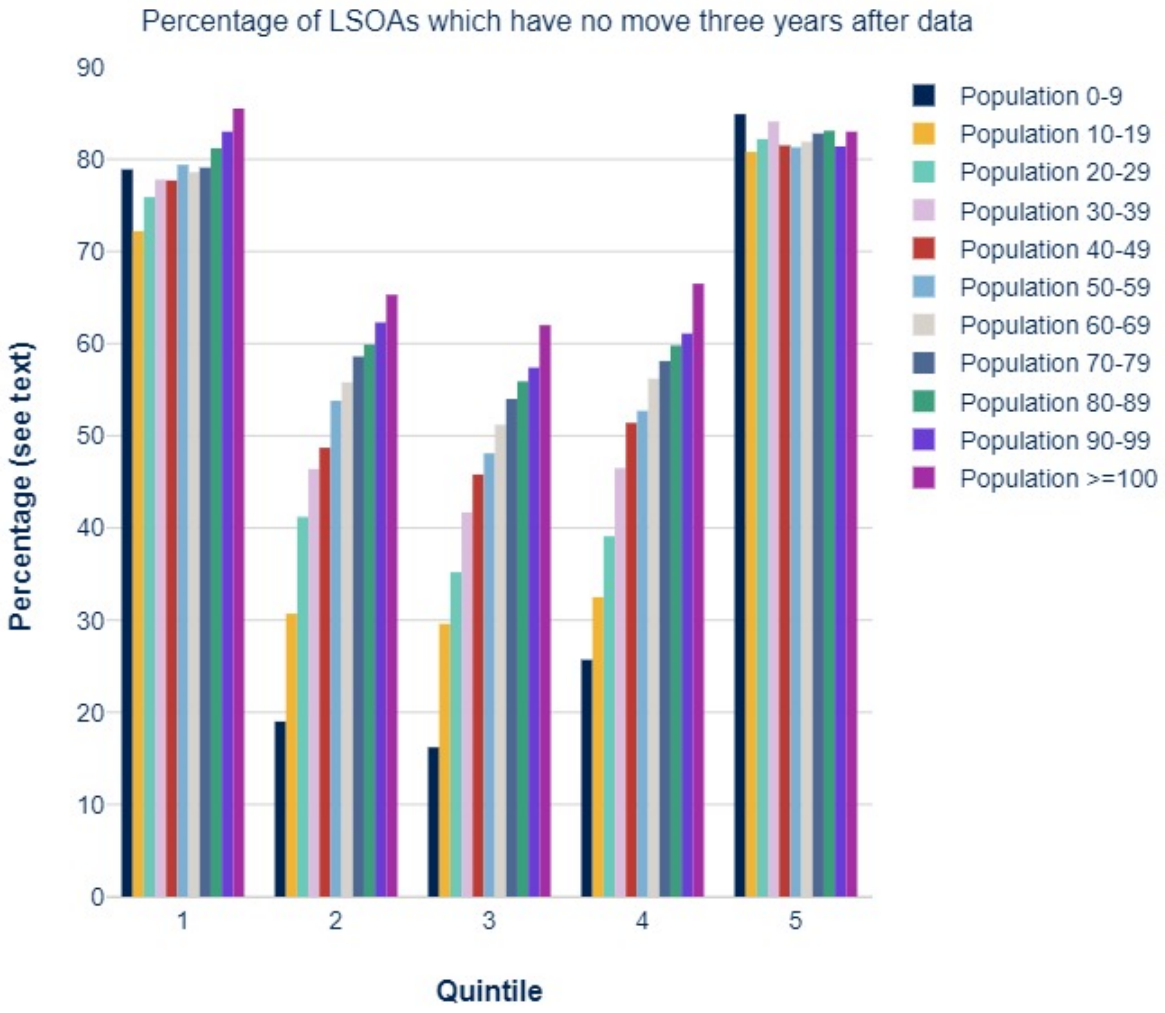


Figure 8: Percentage of LSOAs which move by one quintile three years after data

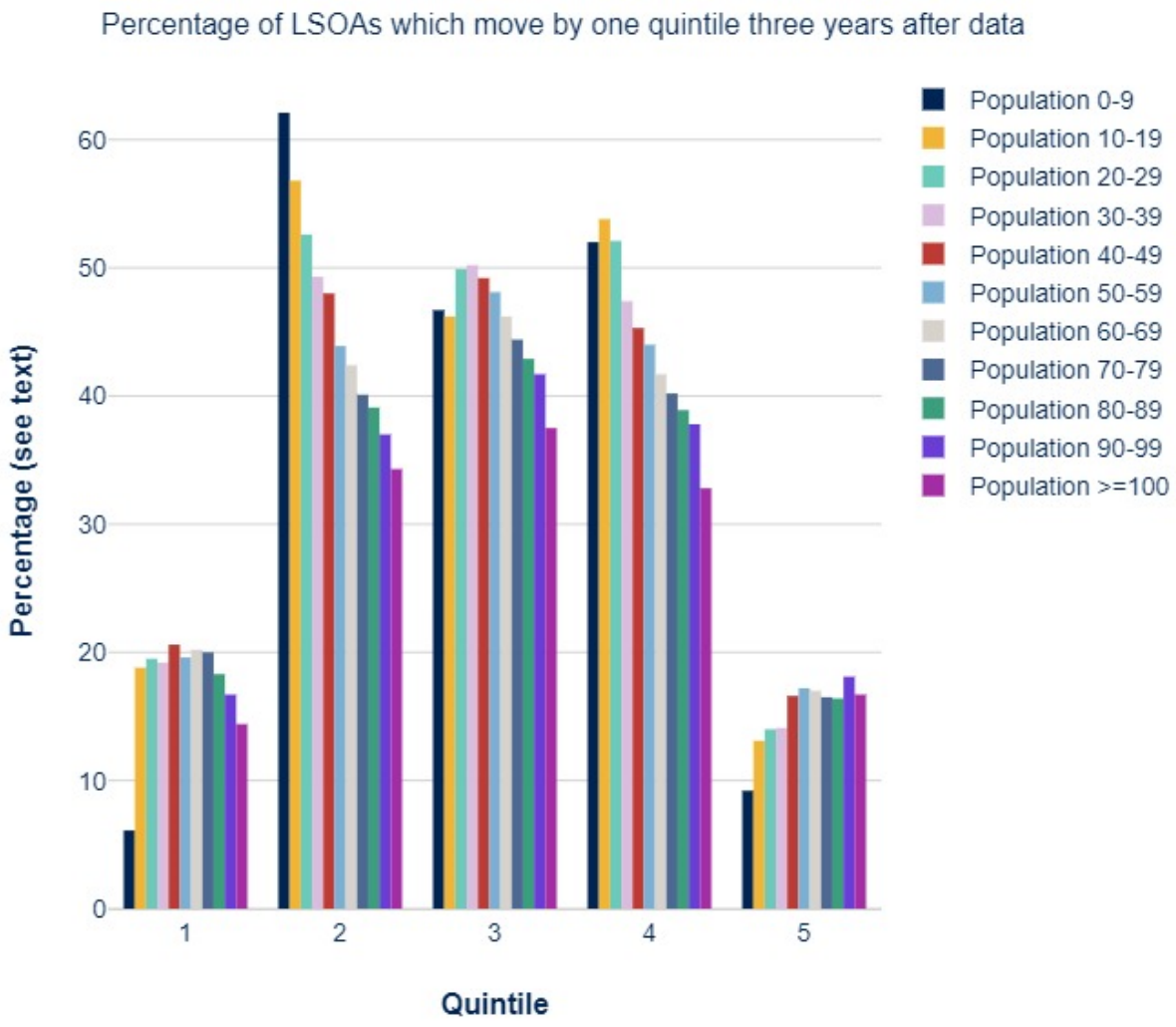
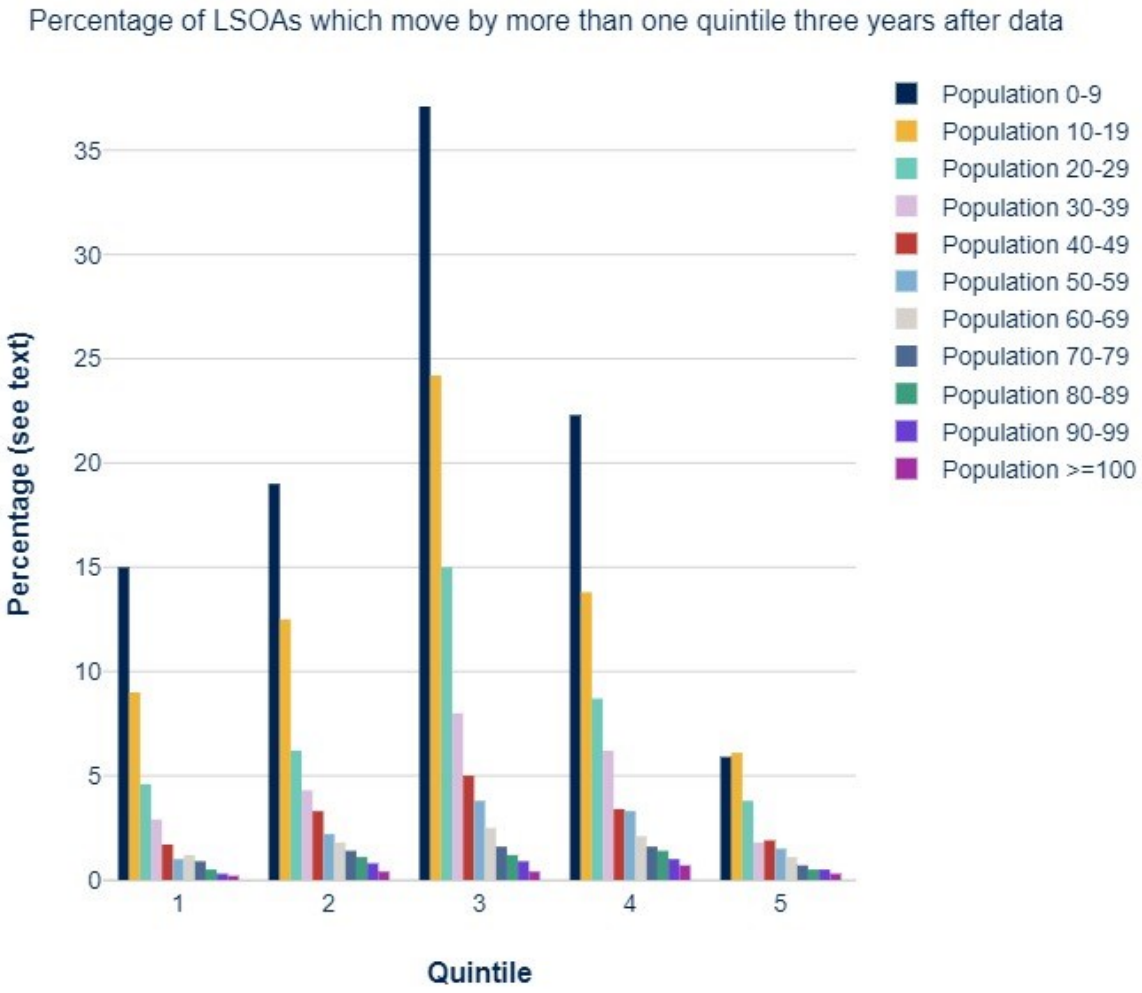


Figure 9: Percentage of LSOAs which move by more than one quintile three years after data



In practice, 1,101 LSOAs of size at least 30 show a large move, (compared to 889 of size at least 50). Simulation suggests that due to random variation, on average 1.1 per cent of LSOAs would show a large move, comprised of 276 (0.87 per cent of total) LSOAs of size at least 50, and a further 71 (0.23 per cent of total) LSOAs size 30-49. The difference between these observed and simulated values suggests the number of LSOAs which show a large move which would not be expected due to random variation, which is now attributed to sensitivity and an underlying change in participation in the LSOA. This is summarised in Table 7.

Table 7: Summary of stability and sensitivity results for a three-year delay

Size of LSOA	Population 30-49	Population at least 50	Population at least 30
Number of large moves in practice	212	889	1,101
Number of large moves expected due to random variation from simulation	71	276	347
Percentage (%) of large moves in practice which are expected due to random variation	33.5	31.0	31.5
Number of large moves expected due to sensitivity	141	613	754
Percentage (%) of large moves in practice which are attributed to sensitivity	66.5	69.0	68.5

Summary

The MSOA classification is very stable and is a good predictor of quintile for an MSOA, even three years after the date of the data. An alternative view of this could be that the classification is not very sensitive to change in the underlying participation rate.

Moving to an LSOA classification introduces more movement, even if suppression below a population of 50 is retained. Reducing the suppression limit to 30 allows even more movement.

Suppose an LSOA is selected from the 0711 classification with suppression below 30. The probability this random LSOA will show a large move is about 1.3 per cent two years later, and about 3.4 per cent three years later.

If the LSOA shows a large move from the 0711 classification:

- two years later: The probability of this large move being due to sensitivity is around 60 per cent. The probability is slightly lower for a population closer to 30 and increases as the base population increases
- three years later: The probability of this large move being due to sensitivity approaches 70 per cent. Again this probability varies with base population.

This work is based on the 0711 classification, taking into account the actual changes in base population for subsequent years.