Office for National Statistics

Article

Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage

Analysis of Statistical Population Dataset version 4.0 2021 using a linkage between Census 2021 and the Demographic Index.

Contact:
Vicky Collison, Sally Mylles, Elizabeth Pereira, Zak Robertson
pop.info@ons.gov.uk
+44 3000 682506

Release date:
28 February 2023

Next release:
To be announced

Table of contents

1 . Main Points

- Linking the Statistical Population Dataset Version 4.0 (SPD v4.0) 2021 to Census 2021 at record level provides unique insights into the quality of SPDs and the effectiveness of its inclusion rules; we will use these insights to make improvements in the future.
- 7.3% of those on Census and Census Coverage Survey were incorrectly excluded from SPD v4.0, while 8.6% in the SPD were incorrectly included.
- Working-age people were more likely to be incorrectly included or incorrectly excluded from the SPD.
- Those in London cosmopolitan and other ethnically diverse metropolitan local authorities were most likely to be incorrectly excluded or incorrectly included.
- The characteristics of those incorrectly included or incorrectly excluded within local authorities can differ, which makes aggregate comparisons less representative in some local authorities.

These are not official statistics and should not be used for decision making. They are estimates from a new methodology different from that currently used to produce official population and migration statistics. They are also based on a sample not representative of the national population, so national-level conclusions should not be made. The information and research in this article should be read alongside the estimates to avoid misinterpretation. These outputs must not be reproduced without this warning.

2 . Background

This research forms part of our population and social statistics transformation programme, which aims to provide the best insights on population, migration and society using a range of data sources. The findings will form part of the evidence base for the National Statistician's Recommendation in 2023 (PDF, 249KB) on the future of population, migration and social statistics in England and Wales.

In this article we directly compare records found in administrative data to those in Census 2021 and the Census Coverage Survey (CCS). This gives us unique insights into the quality of Statistical Population Dataset version 4.0 2021 (SPD v4.0 2021) and the effectiveness of our inclusion rules. We did this with a linkage between Census 2021 and CCS (together referred to as CC) and the Demographic Index (DI) (PDF, 550KB) to determine records that should and should not be in SPD v4.0.

The DI is a composite dataset, built from a range of administrative data sources, providing a solution for the Office for National Statistics (ONS) to work with linked data. The SPD is a subset of the DI, with records included in the SPD if they meet inclusion rules designed to approximate the usually resident population. We compared SPD v4.0 2021 with Census 2021 counts at local authority (LA) and Output Area level.

Comparisons at aggregate level provide an overview of the SPD's performance. However, this does not show what happens at record level where we miss some people we should include and include some people that we should not. This linked analysis using data with identifying information removed helps us understand the types of people that we miss or incorrectly include. For example, the SPD may incorrectly exclude 1,500 people in an LA but also incorrectly include 1,000 people. When compared with census at aggregate level, we see undercoverage of 500 people, because the 1,000 incorrectly included cancel out 1,000 of the incorrectly excluded.

These "incorrect exclusions" and "incorrect inclusions" are therefore distinct from aggregate undercoverage and overcoverage but can be considered undercoverage and overcoverage at a record level. We use the terms "incorrect exclusions" and "incorrect inclusions" to differentiate them from aggregate-level coverage analysis.

Aggregate analysis told us that there are differences in the performance of the SPD by age and sex. The SPD is one of the core sources used in the Dynamic Population Model (DPM), which uses statistical modelling techniques and demographic insights alongside a range of data sources to produce coherent and timely estimates of the population and population change. Understanding the SPD in more detail will help us develop statistical methods to adjust it for bias and improve its accuracy.

Using the linked CC to DI to analyse SPD v4.0 2021, we focused on three groups:

- incorrect exclusions – usual residents that matched between the DI and the CC and that our inclusion rules did not put in the SPD
- incorrect inclusions – non-usual residents and clusters that did not match between the DI and the CC but that our inclusion rules put in the SPD
- correct exclusions – non-usual residents and clusters that did not match between the DI and the CC and that our rules did not put in the SPD

Analysis of these groups enabled us to understand the SPD's coverage issues and the performance of its inclusion rules.

We have used the terms "incorrect" and "correct" exclusions and inclusions in this article. However, the linkage only included those who returned a census form or took part in CCS. Because of this undercoverage, a very small number of these decisions may be erroneous.

We took a 50% sub-sample of the CCS postcodes from the linked CC data (referred to as CCS2). This allowed us to use clerical review in focused areas, increasing the linkage accuracy and preventing missed matches in these postcodes. All analyses in this article focus on these CCS2 areas, but this makes drawing national level conclusions more complex, and we strongly advise against this.

3 . Incorrect exclusions (undercoverage)

There were 512,205 (99.8%) Census 2021 – Census Coverage Survey (CC) records in a CCS2 area that matched to the Demographic Index (DI) and were usual residents. Of these, 7.3% (37,400) were incorrectly excluded from the Statistical Population Dataset version 4.0 2021 (SPD v4.0 2021). This does not include records in CC that were not in the DI. These records should and could be included in the SPD, but our inclusion rules removed them.

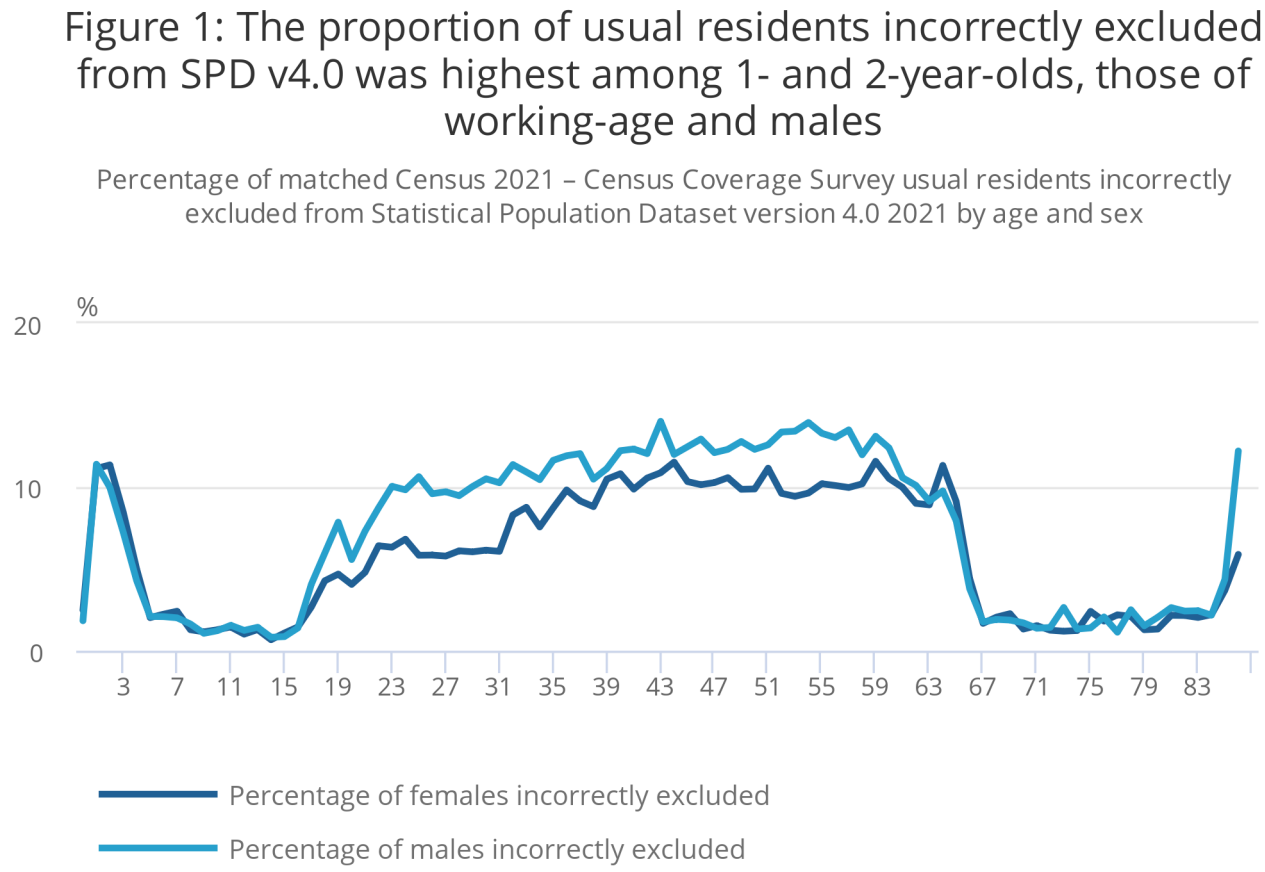
Age and sex

We incorrectly excluded a high proportion (around 11%) of records for those aged between 1 and 2 years from SPD v4.0 2021. This may be because this age group are more likely to be on only one DI data source, and therefore have fewer instances of activity that would lead to their inclusion.

The proportion incorrectly excluded increased from around age 17 years and was highest around age 43 years, before decreasing from the age of 60 years. Around age 18 years, a difference in the proportion of males and females incorrectly excluded also emerges, with males having a greater proportion incorrectly excluded across most ages. This suggests that the inclusion rules inadequately captured working age people, particularly males. This may reflect lower levels of interaction with services that show as activity on data sources at these ages. Male undercoverage is also evident at aggregate level from ages 44 to 66 years, although female undercoverage is greater.

Figure 1: The proportion of usual residents incorrectly excluded from SPD v4.0 was highest among 1- and 2-year-olds, those of working-age and males

Percentage of matched Census 2021 – Census Coverage Survey usual residents incorrectly excluded from Statistical Population Dataset version 4.0 2021 by age and sex



Source: Office for National Statistics - Census 2021

4 . Incorrect inclusions (overcoverage)

In Statistical Population Dataset version 4.0 2021 (SPD v4.0 2021), 518,180 (0.9%) records had an admin data or Census 2021 – Census Coverage Survey (CC) record in a CCS2 area. Of these, 8.4% (43,280) did not match to CC or did match but were a non-usual resident, so were incorrectly included in SPD v4.0.

We removed records for those born after Census Day (21 March 2021) from SPD v4.0 2021 as they should not count as incorrect inclusions. Anyone born after Census Day is not present in CC, but the SPD will include these as it aims to approximate the population at midyear (30 June 2021). We were unable to exclude immigrants that joined the population after this time, so we can attribute some incorrect inclusions to this.

Age and sex

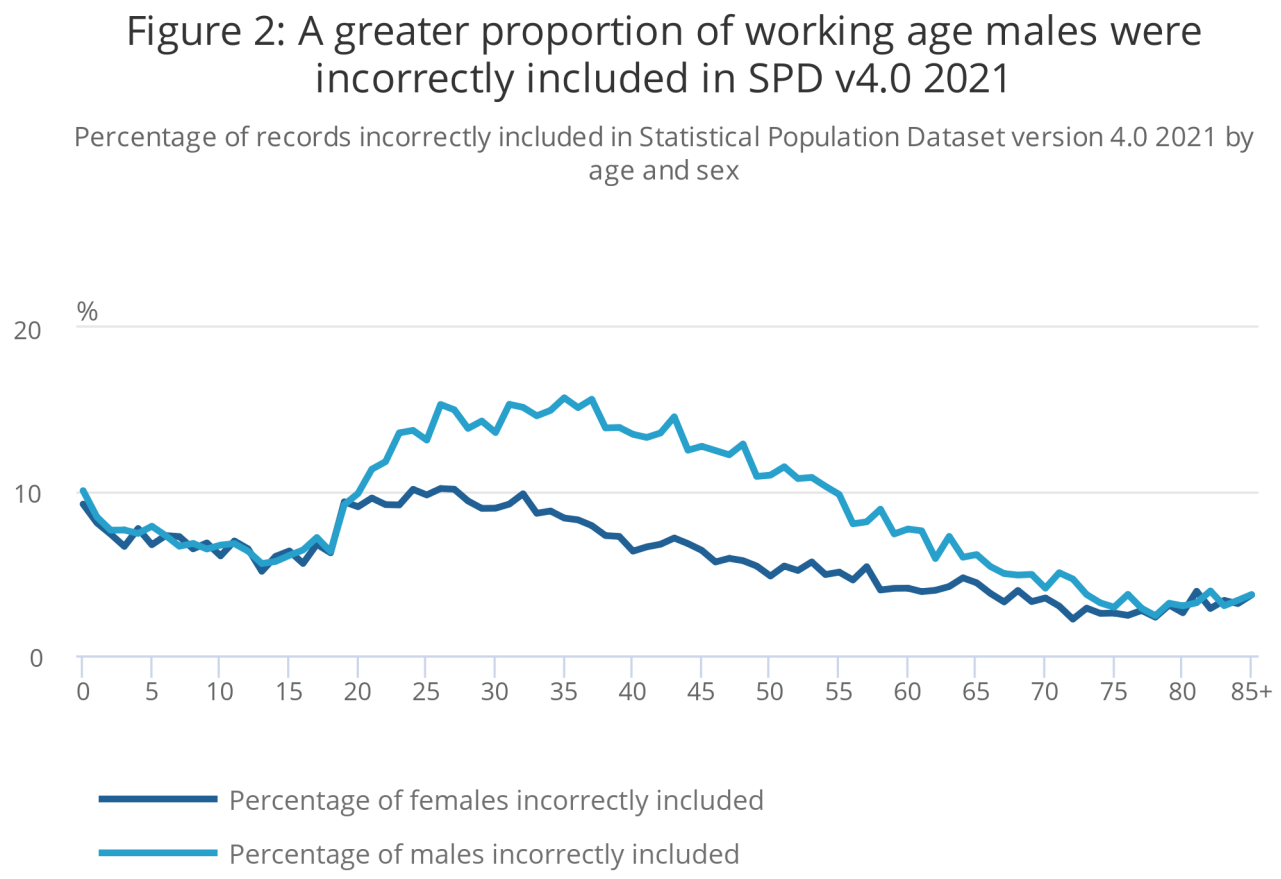
We found an increase in the proportion incorrectly included from age 19 years, particularly for males. This peaked at 15.7% for males aged 35 years before decreasing for later ages. The increase in the proportion incorrectly included is less pronounced for females and peaks around age 27 years (10.1%).

This pattern is broadly similar to those incorrectly excluded from SPD v4.0 2021, but the difference between males and females was more pronounced for incorrect inclusions. This differs from findings at aggregate level, which only show overcoverage from ages 23 to 40 years for males, while females had undercoverage for most working ages. This may reflect the cancelling out of incorrect inclusions and exclusions that can occur at aggregate level.

The similarity between the types of records incorrectly included and excluded from SPD v4.0 2021 reflects that these groups may be less likely to interact with services that lead to them being on or active in one of the administrative sources used to produce the SPD. This could lead to lags in their information being updated when they leave the population (incorrect inclusion). We also use updates to information as an activity indicator for SPD inclusion rules. If people have not recently updated their information, they are more likely to be incorrectly excluded.

Figure 2: A greater proportion of working age males were incorrectly included in SPD v4.0 2021

Percentage of records incorrectly included in Statistical Population Dataset version 4.0 2021 by age and sex



Source: Office for National Statistics - Census 2021

SPD activity

To understand which SPD v4.0 rules contributed to incorrect inclusions in the SPD, we analysed which sources records had activity for. Almost half of records (43.8%, 8,335 records) with activity on only Patient Demographic Service (PDS) were incorrectly included. This may reflect lags on PDS, where people have left the population, but not had their record updated. We incorrectly included a similar number of records (7,335) with activity only on P14 or Tax Credits. However, this only made up 6.3% of records in the sample with activity on only P14 or Tax Credits. As activity on P14 or Tax Credits comes from a single data source, this suggests we were more likely to incorrectly include records showing activity on one of these data sources. Our inclusion rules may therefore need to be refined for records with activity on only these sources.

5 . Correct exclusions

We apply inclusion rules to the Demographic Index (DI) to determine who should be included in the Statistical Population Dataset (SPD). Therefore, it is important to understand where we correctly exclude records from the DI when trying to approximate the usually resident population in the SPD.

There were 26.4% (185,855) of DI clusters not linked to Census 2021 or the Census Coverage Survey (CC), or linked but to non-usual residents in CC. Around 77% (142,595) of these were not on the SPD, so were correctly excluded. The remaining (23%) were incorrect inclusions, which we analysed in the previous section.

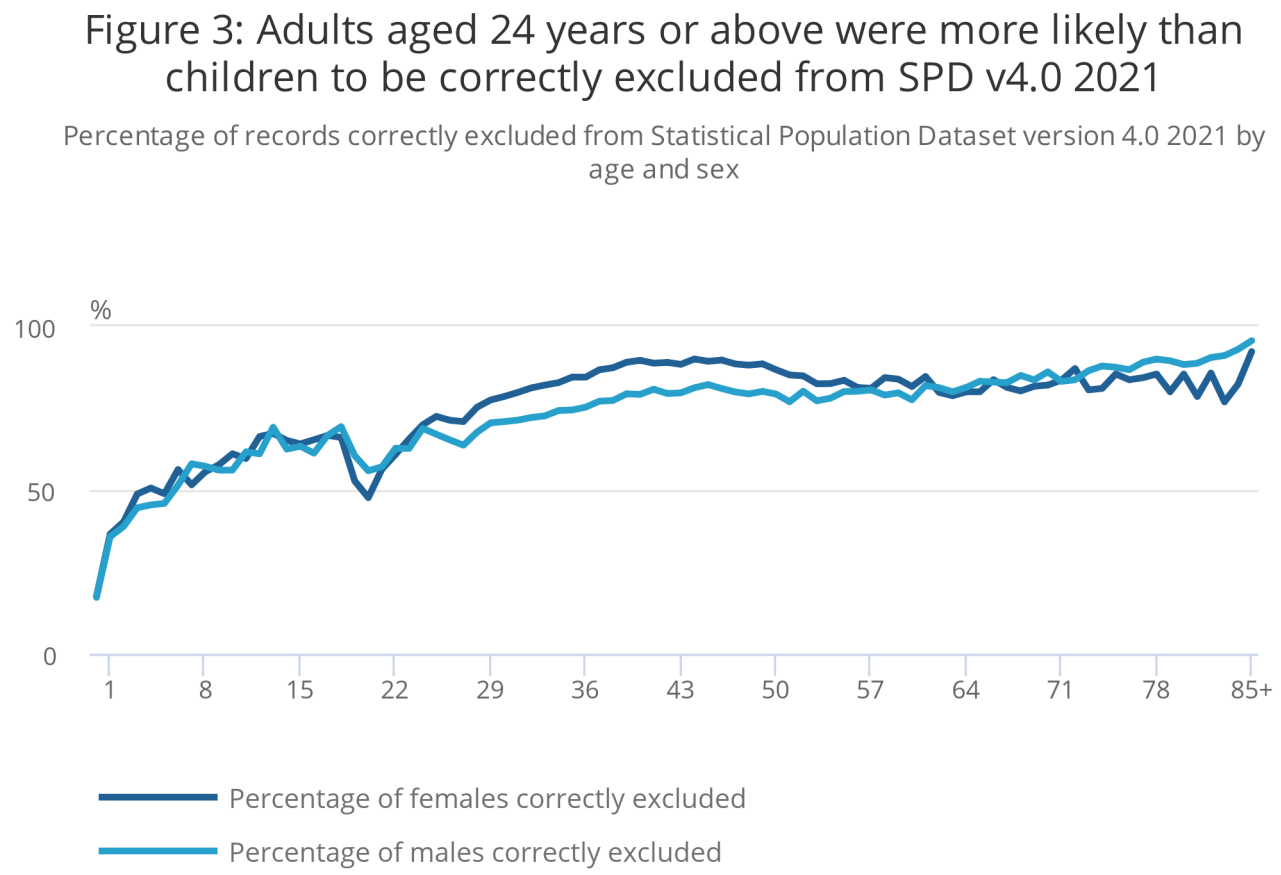
Age and sex

There was a steady increase in the proportion of females correctly excluded from early ages to around age 46 years, where it reaches 89.7%, shown in Figure 3. After this age, the proportion decreased towards retirement age before increasing at older ages. The pattern is similar for males, but the proportion correctly excluded increased more consistently across the ages.

There is a sharp drop in the proportion correctly excluded at ages 19 to 21 years, reaching as low as 47.7% for females and 55.9% for males. This suggests that SPD v4.0 inclusion rules were less effective in correctly excluding people in this age range, reflecting transitions between education and employment.

Figure 3: Adults aged 24 years or above were more likely than children to be correctly excluded from SPD v4.0 2021

Percentage of records correctly excluded from Statistical Population Dataset version 4.0 2021 by age and sex



Source: Office for National Statistics - Census 2021

6 . Geography

For the Statistical Population Dataset version 4.0 2021 (SPD v4.0 2021), most local authorities (LAs) had less than 10% of records incorrectly excluded and incorrectly included (315 and 287 respectively).

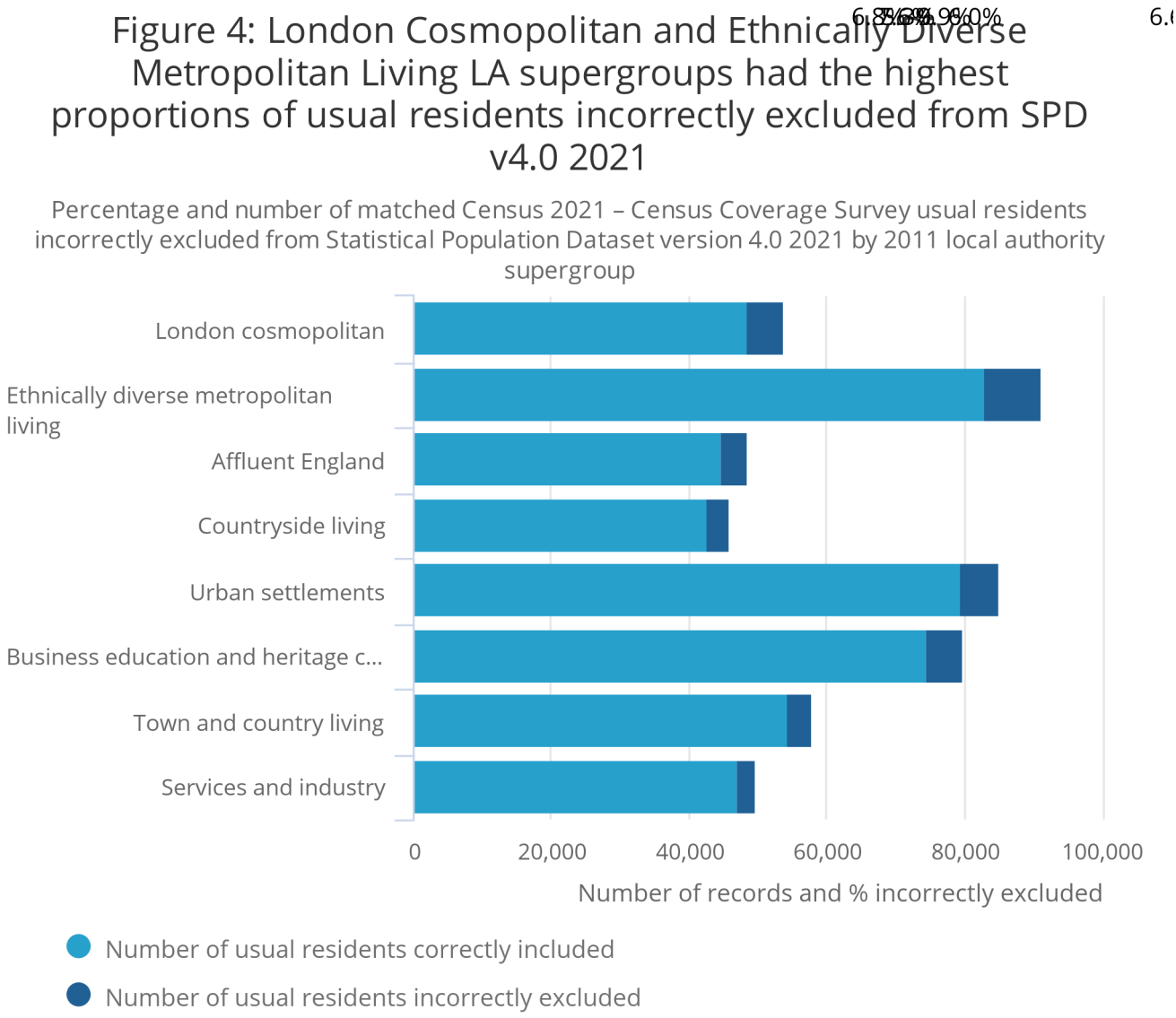
Of the 20 LAs with the greatest proportion incorrectly excluded, six were also in the list of 20 LAs with the greatest proportion incorrectly included. These were Kensington and Chelsea, Westminster, Harrow, Hammersmith and Fulham, City of London and Ealing – all LAs in Greater London. This is expected as these are areas of high population churn.

All of these LAs but Kensington and Chelsea had overcoverage at aggregate level, with SPD estimates for City of London and Westminster more than 5% higher than census estimates. The proportion incorrectly included was also higher than that incorrectly excluded for these LAs. This suggested that in these LAs, incorrect exclusions may cancel out some of the incorrect inclusions, leading to aggregate-level overcoverage.

We grouped LAs using the 2011 Census area classifications supergroups. These supergroups bring together LAs that share common characteristics and provide a way to look at patterns across different types of LAs. The London Cosmopolitan and Ethnically Diverse Metropolitan Living supergroups had the greatest proportion of incorrectly excluded and incorrectly included, shown in Figures 5 and 6. These populations are likely to be more mobile than other types of LA, meaning they may be harder to pick up on the SPD. Lags in updates may mean areas with high mobility have more records in administrative data that are no longer present, leading to greater incorrect inclusion compared with other areas.

Figure 4: London Cosmopolitan and Ethnically Diverse Metropolitan Living LA supergroups had the highest proportions of usual residents incorrectly excluded from SPD v4.0 2021

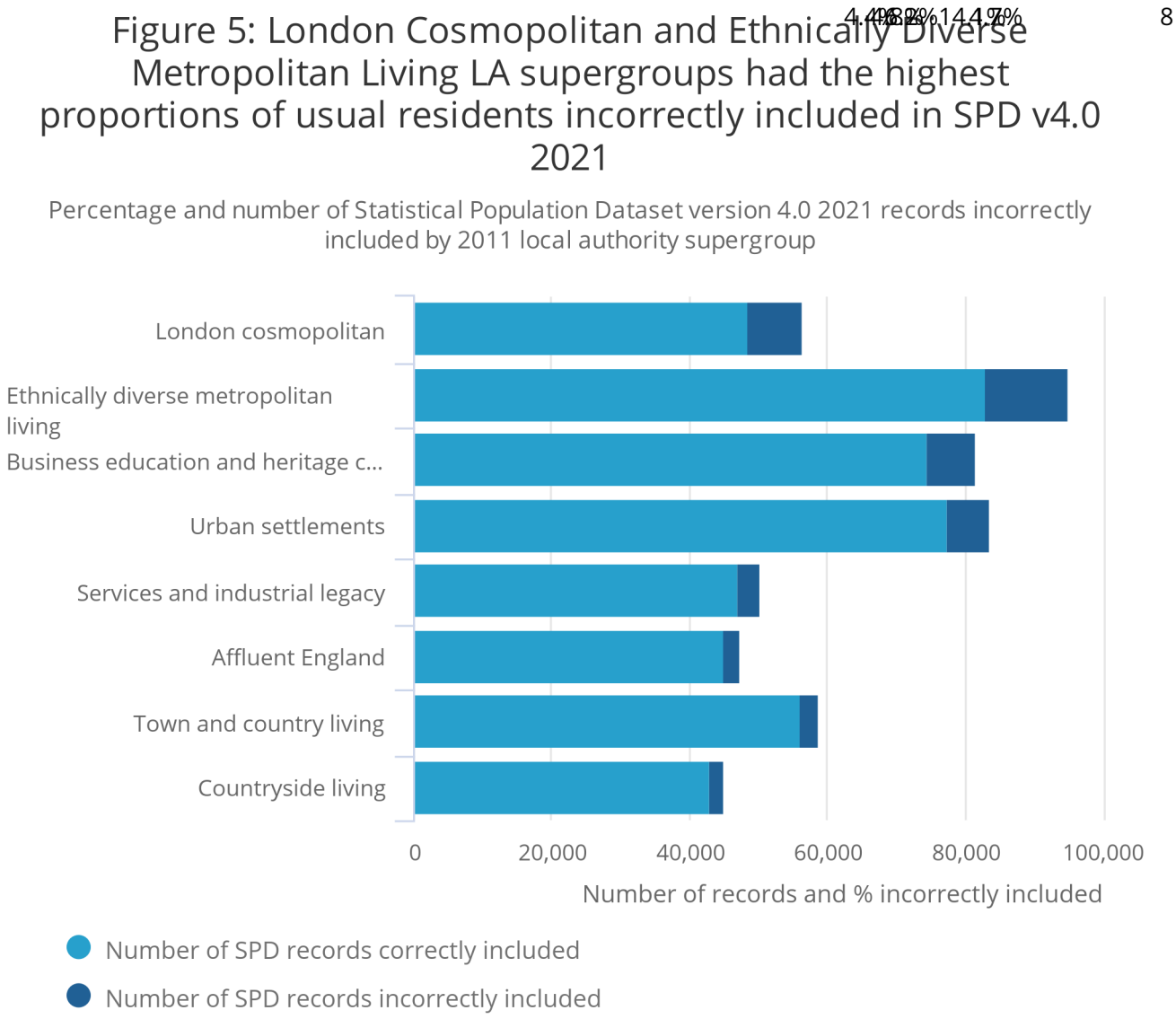
Percentage and number of matched Census 2021 – Census Coverage Survey usual residents incorrectly excluded from Statistical Population Dataset version 4.0 2021 by 2011 local authority supergroup



Source: Office for National Statistics - Census 2021

Figure 5: London Cosmopolitan and Ethnically Diverse Metropolitan Living LA supergroups had the highest proportions of usual residents incorrectly included in SPD v4.0 2021

Percentage and number of Statistical Population Dataset version 4.0 2021 records incorrectly included by 2011 local authority supergroup



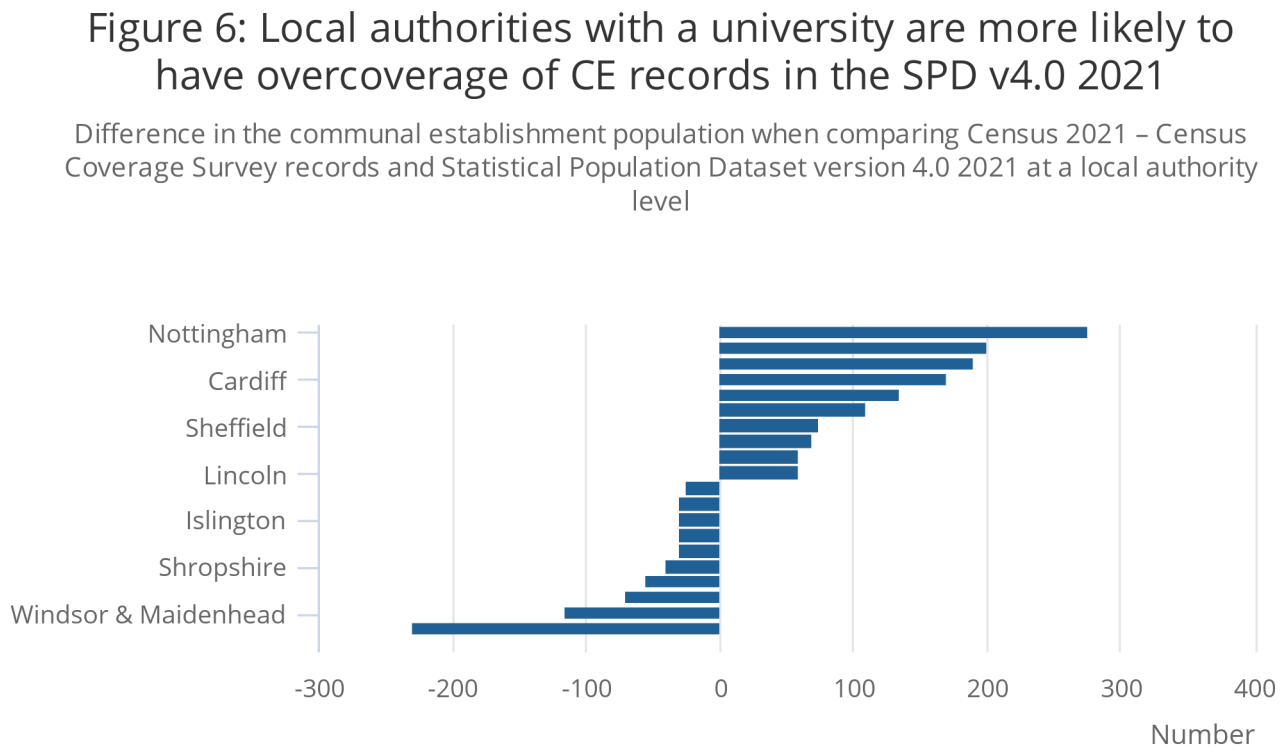
Source: Office for National Statistics - Census 2021

7 . Communal establishments

Previous analysis of the Statistical Population Dataset (SPD) by Output Area (OA) showed a relationship between OAs with large differences to the 2011 Census and OAs containing communal establishments (CEs). The 2021 SPD has challenges at an aggregate level, however, looking at the record level can inform how we improve CE estimation in the future.

Figure 6: Local authorities with a university are more likely to have overcoverage of CE records in the SPD v4.0 2021

Difference in the communal establishment population when comparing Census 2021 – Census Coverage Survey records and Statistical Population Dataset version 4.0 2021 at a local authority level



Source: Office for National Statistics - Census 2021

Our current method for assigning record clusters to CEs had substantial overcoverage of people living in halls of residence according to admin data. This likely explains why the 10 local authorities (LAs) with the largest admin data overcoverage contain one or more university.

Hart is the LA with the largest administrative data undercoverage, as there are 230 more records in the Census – Census Coverage Survey (CC). Hart has a military base found on the census. However, we do not have administrative data records for this address.

Windsor and Maidenhead has the second-largest admin data undercount. Most of these records come from those assigned to boarding schools in CC but not in the SPD. Boarding school pupils are a possible area of undercoverage in the SPD as they are not included in the English and Welsh School Censuses (ESC and WSC).

Looking at record level, most people who appear in CC data within a CE are linked to an administrative data cluster. Less than 0.5% of CC CE residents were not linked to the administrative data. However, only 33% of records found in both the SPD version 4.0 2021 (SPD v4.0 2021) and CC are assigned to the CE population in both data sources. This means in the SPD v4.0 2021 we are incorrectly assigning records to households when they should be in a CE.

The rules that create the SPD v4.0 successfully removed Demographic Index (DI) records, which were not linked to the CC. Only 1.3% of the unlinked administrative data records assigned to a CE were incorrectly included in the SPD. The SPD rules do, however, remove 4% of correct CE matches.

8 . Case study: Harrow

Harrow has both a high proportion of incorrect exclusions (11.4%) and incorrect inclusions (13.5%), appearing in both the top 20 local authorities (LAs) with the greatest proportion incorrectly excluded and incorrectly included. Compared with census at aggregate level, Harrow appears to have good coverage; around 1% higher than Census 2021.

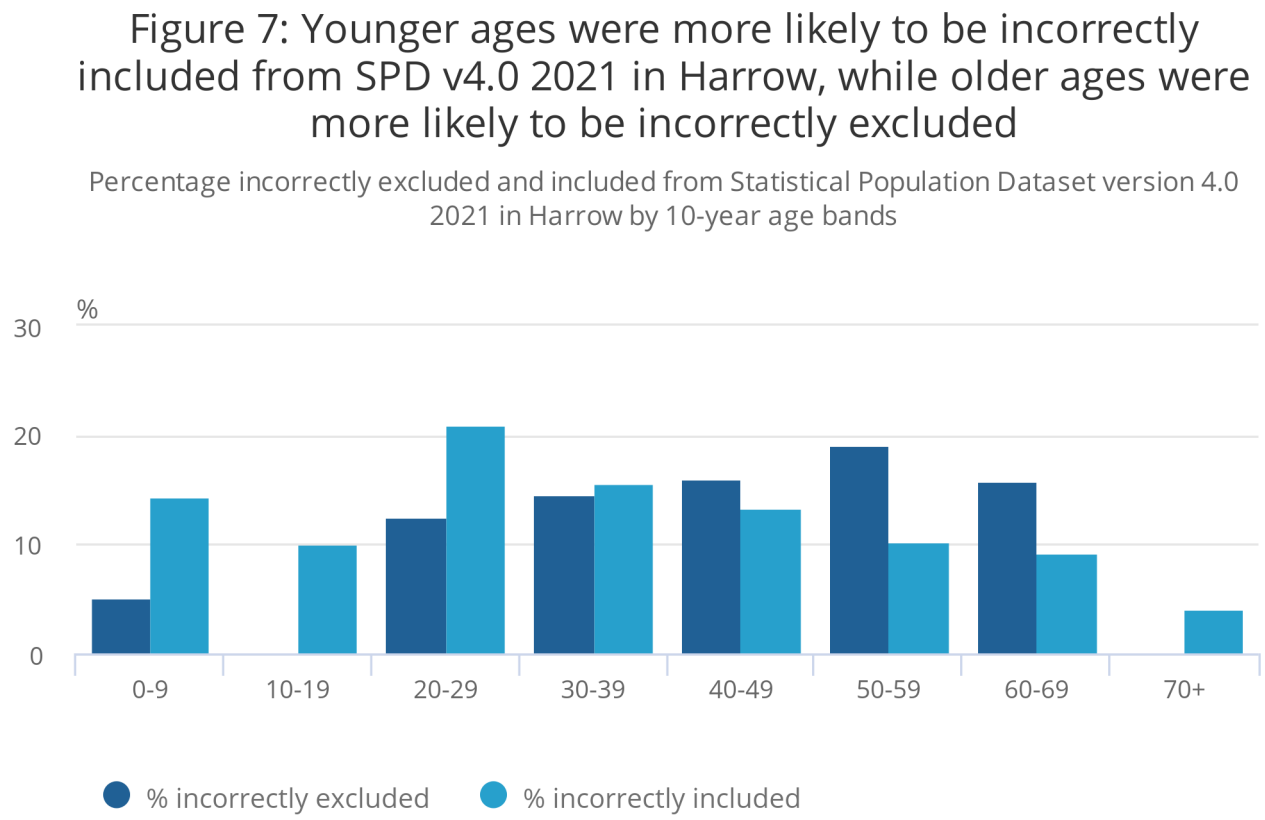
For the purposes of aggregate statistics, having high incorrect inclusions and exclusions is less of an issue when records with the same characteristics are incorrectly included and excluded. This is because the effects at record level cancel each other out at aggregate level, so the characteristics of the records in that LA will look broadly correct.

Figure 6 shows a difference in the age profile of the 305 incorrectly excluded and 385 incorrectly included records in Harrow. Younger people were more likely to be incorrectly included, with 20- to 29-year-olds having the highest proportion (20.9%). Those aged 20 to 29 years appeared on up to three Demographic Index (DI) sources in greater number than in younger or older age bands. This may increase their likelihood of incorrect inclusion in Statistical Population Dataset version 4.0 (SPD v4.0) because of activity on multiple sources before leaving the population.

In contrast, older people were more likely to be incorrectly excluded, with 50- to 59-year-olds having the highest proportion (19.1%). Groups in this age band may have reduced interaction with services required for inclusion in SPD v4.0 because of early retirement or not needing to work. The self-employed may also be excluded as self-assessment data are not included in SPD v4.0. This suggests that in Harrow, records with different characteristics are incorrectly included and excluded, which may affect analysis at aggregate level. However, because of the size of the present sample and the bias in analysis restricted to CCS2 areas, this cannot be determined for certain.

Figure 7: Younger ages were more likely to be incorrectly included from SPD v4.0 2021 in Harrow, while older ages were more likely to be incorrectly excluded

Percentage incorrectly excluded and included from Statistical Population Dataset version 4.0 2021 in Harrow by 10-year age bands



Source: Office for National Statistics - Census 2021

9 . Glossary

Administrative data

Collections of data maintained for administrative reasons, for example, registrations, transactions, or record-keeping. They are used for operational purposes and their statistical use is secondary. These sources are typically managed by other government bodies.

Data linkage

Data linkage is the process of joining together records that relate to the same entity, such as a person or business.

Correct exclusions

Those that the linkage suggests should not have been in the Statistical Population Dataset (SPD) because they were non-usual residents or did not match between the Demographic Index (DI) and the Census 2021 - Census Coverage Survey (CC) and that our rules did not put in the SPD.

Incorrect exclusions

Those that the linkage suggests should have been in the SPD because they were usual residents that matched between the DI and the CC but were not included in the SPD.

Incorrect inclusions

Those that the linkage suggests should not have been in the SPD because they were non-usual residents or did not match between the DI and the CC but that our inclusion rules put in the SPD.

Usual residents

A usual resident of the UK is anyone who, on 21 March 2021, is in the UK and has stayed, or intends to stay, in the UK for 12 months or more or has a permanent UK address and is outside the UK and intends to be outside the UK for less than 12 months.

P14

Form P14 is a return for an individual taxpayer. An employer must complete a form P14 for each employee where they were required to complete a P11 deductions working sheet during the year.

10 . Data sources and quality

Census

Response to the 2021 Census was very high – around 97%. The quality of the census estimates was considered very high after applying a coverage adjustment process that accounted for non-response and incorrect responses. However, since our analysis for Census Coverage Survey (CCS2) required the matching of individual records, we could only include those who actually responded to census or CCS. In addition,CCS areas, and by extension the CCS2, typically did not include large communal establishments (CEs) (PDF, 208KB) as they are a subject to a separate estimation process. Because the CCS sample is designed to target areas of low census response, and both census and CCS will be subject to non-response bias, those people included in the CCS2 are not representative of the population of England and Wales, which makes generalisation to the whole population difficult.

Census Coverage Survey (CCS)

The 2021 Census took place on 21 March 2021 to capture information on every household in England and Wales. The 2021 CCS began data collection eight weeks later and sampled 1.45% of the postcodes in England and Wales. Postcodes were sampled disproportionately (DOCX, 1.28MB), based on expected census response rates. The data, gathered independently of the census, allowed us to assess the coverage of the Census as well as estimate the true population of England and Wales. The linkage between the census and CCS (CC) was designed to be of an extremely high standard, reducing false and missed matches as much as possible.

SPD v4.0 2021

The Statistical Population Dataset version 4.0 2021 (SPD v4.0 2021) was built using the Demographic Index (DI) V2.1, whereas the version of the Demographic Index linked to Census 2021 – CCS was V2.0. There are small differences between these versions of the DI, such as the later version including Higher Education Statistics Agency (HESA) and CIS data for 2021. However, over 99% of those in SPD v4.0 2021 were also present in V2.0 of the DI. Those that were not present were not included in this analysis.

Administrative data

- Where possible, we have used administrative data that covers 2021.
- Administrative data are collected continuously throughout the year and a snapshot taken at specific times. For the data used in this article, none of the extracts were taken on Census Day (21 March 2021) so there is some time-lag related error possible.
- Where possible we have addressed this (for example, removing babies born after Census Day from the analysis).
- Analysis did not use CIS data to ensure compliance with data sharing agreements held.
- More on the administrative data sources used in this article, and how they were used to support the census process, can be found in the Administrative data used in Census report.

11 . Future developments

This analysis has enabled us to understand the characteristics of records incorrectly included and excluded from Statistical Population Dataset version 4.0 (SPD v4.0), as well as those the SPD rules correctly excluded. This has provided additional insights into analysis of the SPD at aggregate level, showing that similar types of people tend to be incorrectly included and excluded at a national level, although this can differ within specific local authorities.

Our work to produce admin-based population estimates (ABPEs) from the Dynamic Population Model (DPM) has demonstrated the need to put in place a robust coverage adjustment method using a coverage survey to support delivering estimates to a quality that meets user needs. This work provides valuable insight into incorrect inclusions and exclusions that can be used in the design of the coverage adjustment method.

We plan to do further work to understand those incorrectly included and excluded, as well as those correctly excluded. We also plan to explore those correctly included. This further research will help us test and refine the SPD's inclusion rules and the data sources it uses.

12 . Cite this article

Office for National Statistics (ONS), released 28 February 2023, ONS website, article, Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census – Demographic Index linkage.