Office for National Statistics

Article

Understanding quality of linked administrative data sources in England and Wales, using the 2021 Census – Demographic Index linkage

Analysis of linkage between the Demographic Index and linked census and Census Coverage Survey to understand the quality of administrative data sources in England and Wales.

Contact:
Ayesha Barnes, Owen Maynard, Sally Mylles, Chloe Pearce, Elizabeth Pereira and Zak Robertson
pop.info@ons.gov.uk
+44 3000 682506

Release date:
1 March 2023

Next release:
To be announced

Table of contents

1 . Main points

- Less than 1% of the 463,000 Census-Census Coverage Survey (CC) respondents in Census Coverage Survey 2 (CCS2) areas did not link to the administrative data.

- The CC respondents most likely to not be on the Demographic Index (DI) were males aged 20 to 39 years and migrants.

- Using administrative data, we could assign 95% of linked respondents to the same local authority (LA) as their CC LA.

- Most communal establishment residents found in the DI and CC dataset were only located in a communal establishment on one or the other, not both.

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. 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 recommendations on the future of the population statistics system in 2023 (PDF, 249KB), including migration and social statistics in England and Wales.

This report analyses the results of Linkage between the Demographic Index (DI) and linked Census 2021 and Census Coverage Survey (CCS) datasets (CC) (PDF, 523KB). The DI is built from a range of administrative data sources to provide a solution for the Office for National Statistics (ONS) to work with linked data. The administrative data sources, found in our Data source overviews, from the DI that have been used in this linkage are:

- NHS's Personal Demographics Service (PDS)
- Department for Education's English School Census (ESC)
- Welsh Government's Welsh School Census (WSC)
- Higher Education Statistics Authority student data (HESA)
- Department for Work and Pensions' Customer Information System (CIS)

Census 2021 took place on 21 March 2021, to collect information on the population of England and Wales. The 2021 CCS began data collection eight weeks later, as described in our Coverage estimation for Census 2021 in England and Wales. It sampled 1.45% of the postcodes in England and Wales. Postcodes were sampled disproportionately, based on the Hard to Count index for the 2021 Census (DOCX, 1.24MB). 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 Census 2021 and CCS (together referred to as CC) was designed to be of an extremely high standard, reducing false and missed matches as much as possible. More information on this can be found in our Linkage methods for Census 2021 in England and Wales methodology. The combined dataset is referred to as CC throughout.

Following automatic linkage between the DI and CC data, a 50% sub-sample of the CCS postcodes was taken (CCS2), as described in the Linkage project between the 2021 Census and Census Coverage Survey to the Demographic Index (PDF, 523KB). This allowed clerical review to be used in focussed areas, increasing the accuracy of the linkage and greatly reducing the likelihood of missed matches in these postcodes. The CCS2 areas included around 463,000 CC respondents, more than 99% of whom were usual residents in England and Wales. All analyses in this paper focus on these CCS2 areas. However, this does make drawing national level conclusions more complex, as the data is not necessarily representative of the whole population. We would strongly advise against this.

Analysing the linked CC data allows us to achieve numerous aims, such as:

- developing further understanding of the Statistical quality of the DI (PDF, 549KB)
- better informed decisions in the construction of the Statistical Population Datasets (SPD)
- addressing the need for a robust coverage adjustment method for the SPD, as shown in SPD Estimation Options (PDF, 1.3MB), to support its use in the dynamic population model[MO22] (DPM)
- better understanding of how well we capture data for communal establishments (CE)
- better understanding of how well administrative data captures detailed population characteristics
- developing better linkage projects within ONS in the future

The research questions have used one, or more, of the populations outlined in the following:

- DI residuals: records that were not linked to CC and only appear on the DI
- DI to CC linked usual residents: linked records, considered to be usual residents on the census
- DI to CC linked non-usual residents: linked records, not considered to be usual residents on the census
- CC residuals: records that were not linked to the DI and only appear on the census, CCS, or both

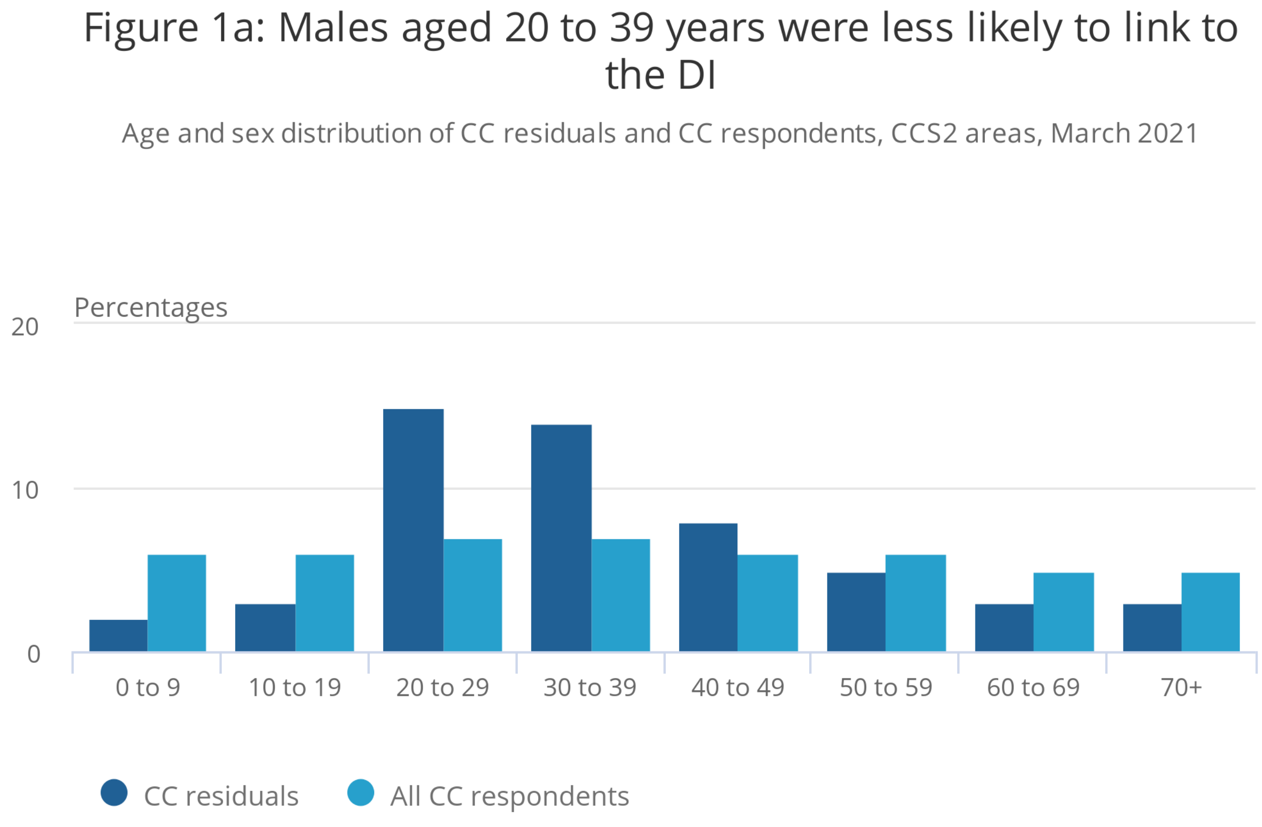
3 . Respondents on the Census-Census Coverage Survey (CCS) in CCS2 areas that did not link to the Demographic Index

Less than 1% of Census-CCS (CC) respondents in CCS2 areas did not link to the Demographic Index (DI), suggesting the DI captures the population well. Comparing these unlinked respondents, referred to as CC residuals, to all CC respondents in CCS2 areas allows us to understand if certain characteristics were more prevalent in the residuals. The small sample size of CC residuals means findings should be interpreted with caution.

Figure 1a shows the age and sex distribution, in which males aged 20 to 39 years comprised a higher percentage of CC residuals (29%) than all CC respondents (14%). Figure 1b suggests that they may not interact with the services that use the administrative systems as much as females.

Figure 1a: Males aged 20 to 39 years were less likely to link to the DI

Age and sex distribution of CC residuals and CC respondents, CCS2 areas, March 2021



Source: Office for National Statistics

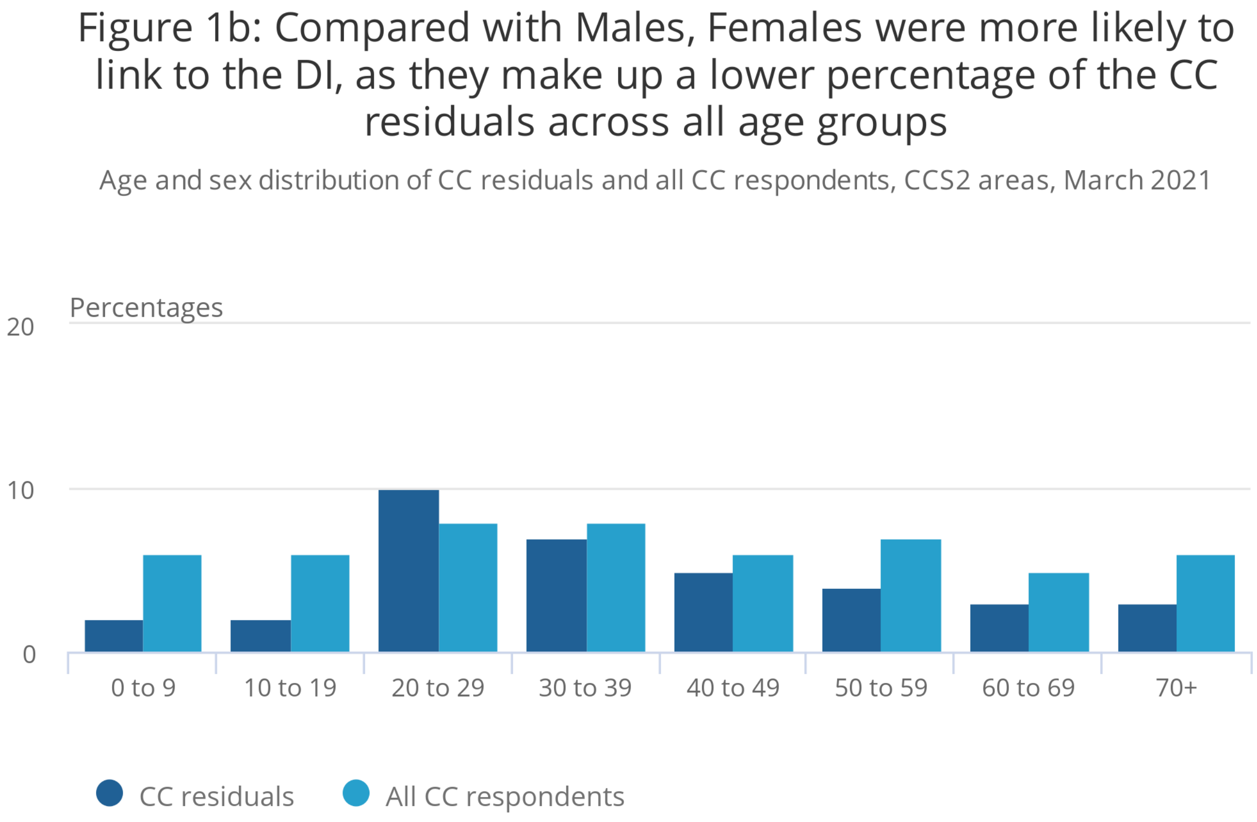
Notes:

1. Percentages may not sum to 100 because of rounding.

2. There were many small counts which required suppression or grouping in line with disclosure control regulations.

Figure 1b: Compared with Males, Females were more likely to link to the DI, as they make up a lower percentage of the CC residuals across all age groups

Age and sex distribution of CC residuals and all CC respondents, CCS2 areas, March 2021



Source: Office for National Statistics

Notes:

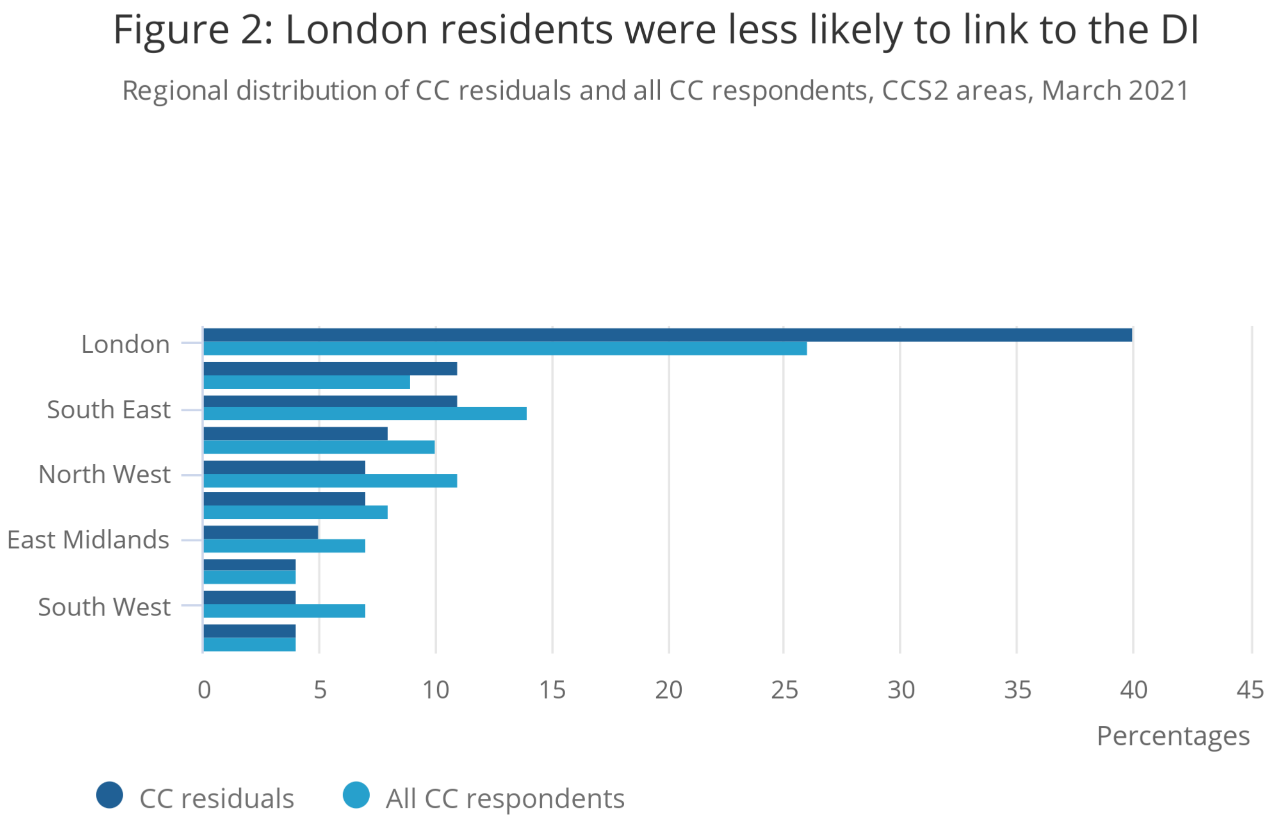
1. Percentages may not sum to 100 because of rounding.

2. There were many small counts which required suppression or grouping in line with disclosure control regulations.

Regional analysis in Figure 2 showed that the percentage of respondents based in London was higher for CC residuals (38%) than all CC respondents (26%). This suggests we may not capture some London residents as accurately on the DI, possibly because this population is more mobile than other regions. This is consistent with our SPD aggregate analysis in our Developing Statistical Population Datasets, England and Wales: 2021 article and Understanding quality of Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage article. This observation is based on a sample of the population, so we advise against applying these findings to the whole population.

Figure 2: London residents were less likely to link to the DI

Regional distribution of CC residuals and all CC respondents, CCS2 areas, March 2021



Source: Office for National Statistics

Notes:

1. Percentages may not sum to 100 because of rounding.

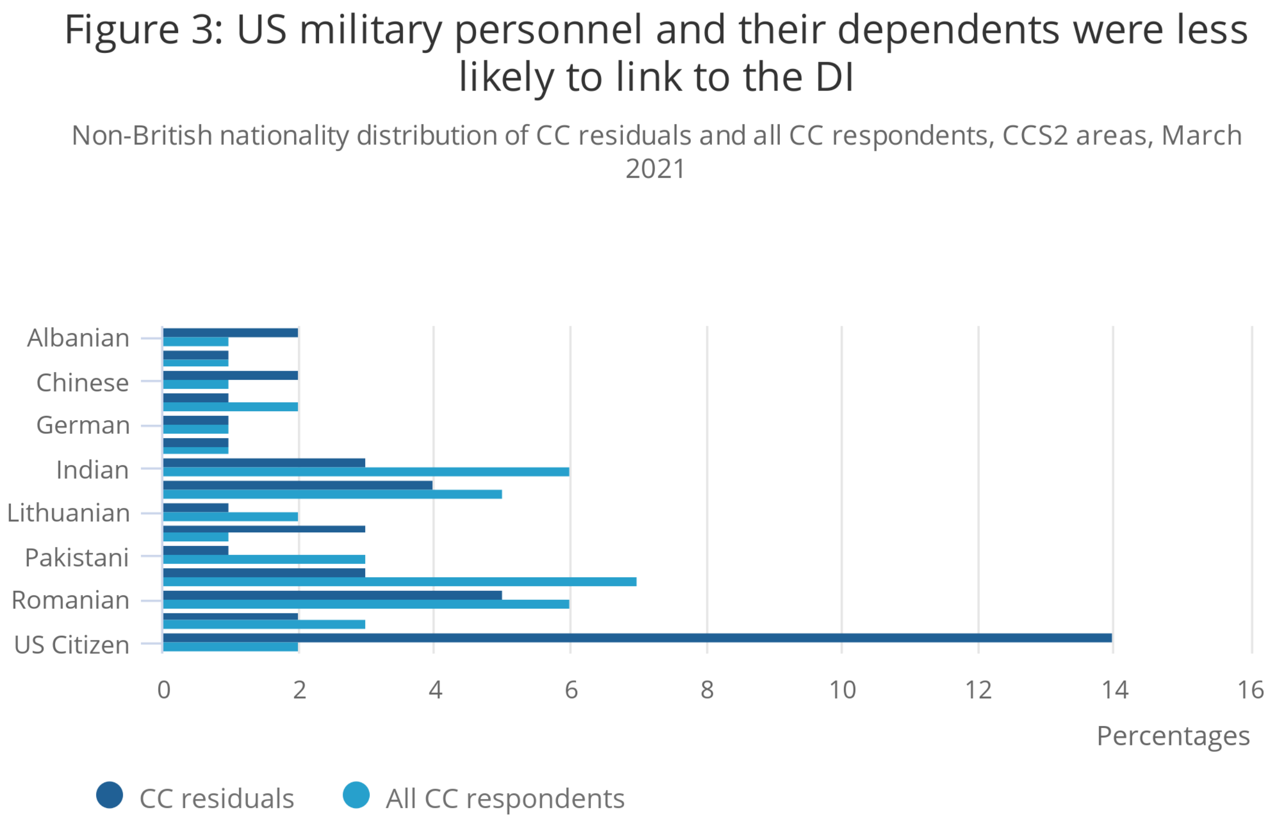
Analyses of country of birth, nationality, and main language variables indicated that a higher percentage of unlinked respondents were migrants, when compared with all CC respondents.

The country of birth breakdown showed 52% of CC residuals had a country of birth outside the UK, compared with 23% of all CC respondents. This group of non-UK born respondents could include those who have recently arrived in the country and have not interacted with services yet. It could also include short-term migrants, who are less likely to interact with services as they are not long-term residents, so may not be included on the DI. Approximately 7% of CC residuals were flagged as not usually resident, compared with less than 1% of all CC respondents.

Breakdowns by nationality in Figure 3 showed a larger percentage of US citizens in the CC residuals than in all CC respondents. Of these, 64% had a postcode in West Suffolk or Kings Lynn and West Norfolk, locations of US air bases. We would not expect this population to appear on the DI. Foreign military personnel, and their dependents, typically use alternative healthcare, income, and schooling systems. Adjustments will need to be made for this population, as part of the coverage adjustment strategy in the DPM.

Figure 3: US military personnel and their dependents were less likely to link to the DI

Non-British nationality distribution of CC residuals and all CC respondents, CCS2 areas, March 2021



Source: Office for National Statistics

Notes:

1. Percentages may not sum to 100 because of rounding.

2. This is a census-only variable so CCS-only responses have been removed.

3. There were many small counts which required suppression in line with disclosure control regulations.

The census main language variable showed 22% of CC residuals did not speak English as their main language, compared with 10% of all CC respondents. Not having English as a main language may be a barrier to accessing services whose administrative systems generate the data used to construct the DI.

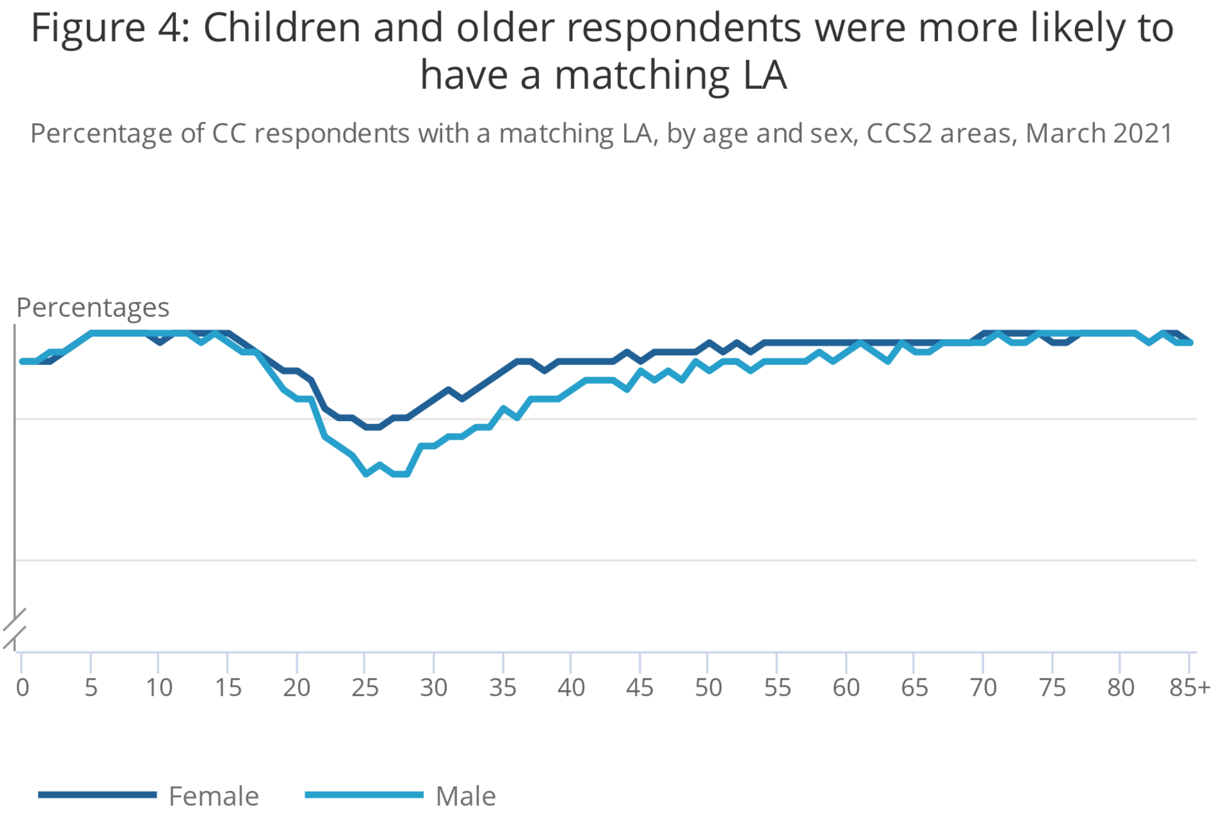
4 . Percentage of linked respondents with matching address information on the Demographic Index (DI), when compared with the census or Census Coverage Survey (CCS)

To understand the accuracy of the address information we can assign to Census-CCS (CC) respondents using the DI, we compared the CC address local authorities (LA) to the DI address LAs. We found that 95% of linked respondents had at least one LA that matched.

Children and older respondents were more likely to have a matching LA, and those aged 22 to 35 years were least likely to. Figure 4 shows that, from 19 to 60 years old, men were less likely than women to have a matching LA . This may be because children, working-aged women, and older people are more likely to interact with services that collect administrative data, such as the NHS and schools. Households with children and older people may also move less frequently than other groups, such as university students.

Figure 4: Children and older respondents were more likely to have a matching LA

Percentage of CC respondents with a matching LA, by age and sex, CCS2 areas, March 2021

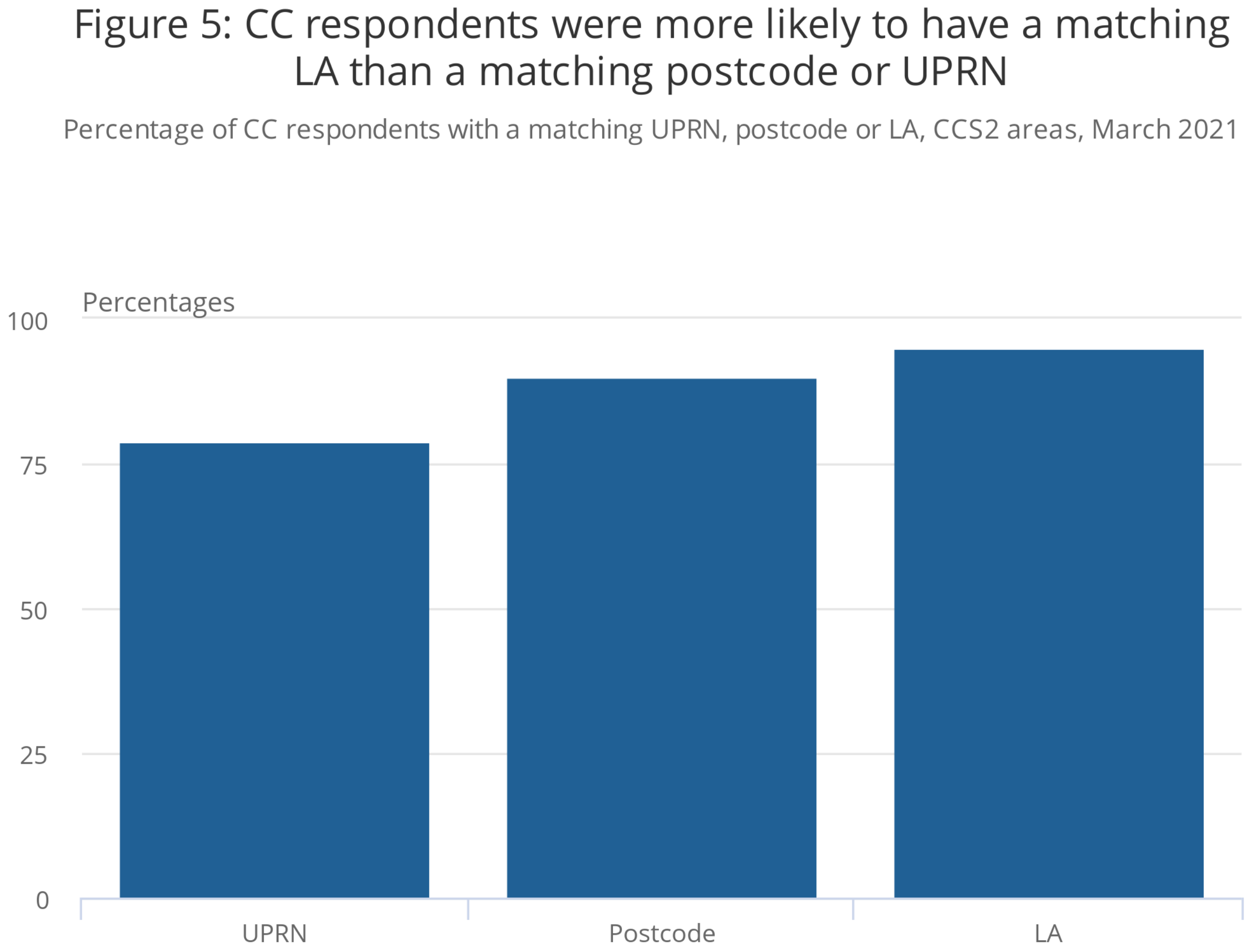


Source: Office for National Statistics

We also looked at the percentage of respondents who had a matching postcode and Unique Property Reference Number (UPRN). Information on UPRNs can be found in government Identifying property and street information guidance. Figure 5 shows that 90% had a matching postcode and 79% had a matching UPRN. There was a higher percentage of respondents with a matching LA, because this is a larger geographic area than postcode or UPRN. As such, even if the postcode or UPRN does not match across the DI and the CC, the LA still can.

Figure 5: CC respondents were more likely to have a matching LA than a matching postcode or UPRN

Percentage of CC respondents with a matching UPRN, postcode or LA, CCS2 areas, March 2021



Source: Office for National Statistics

Notes:

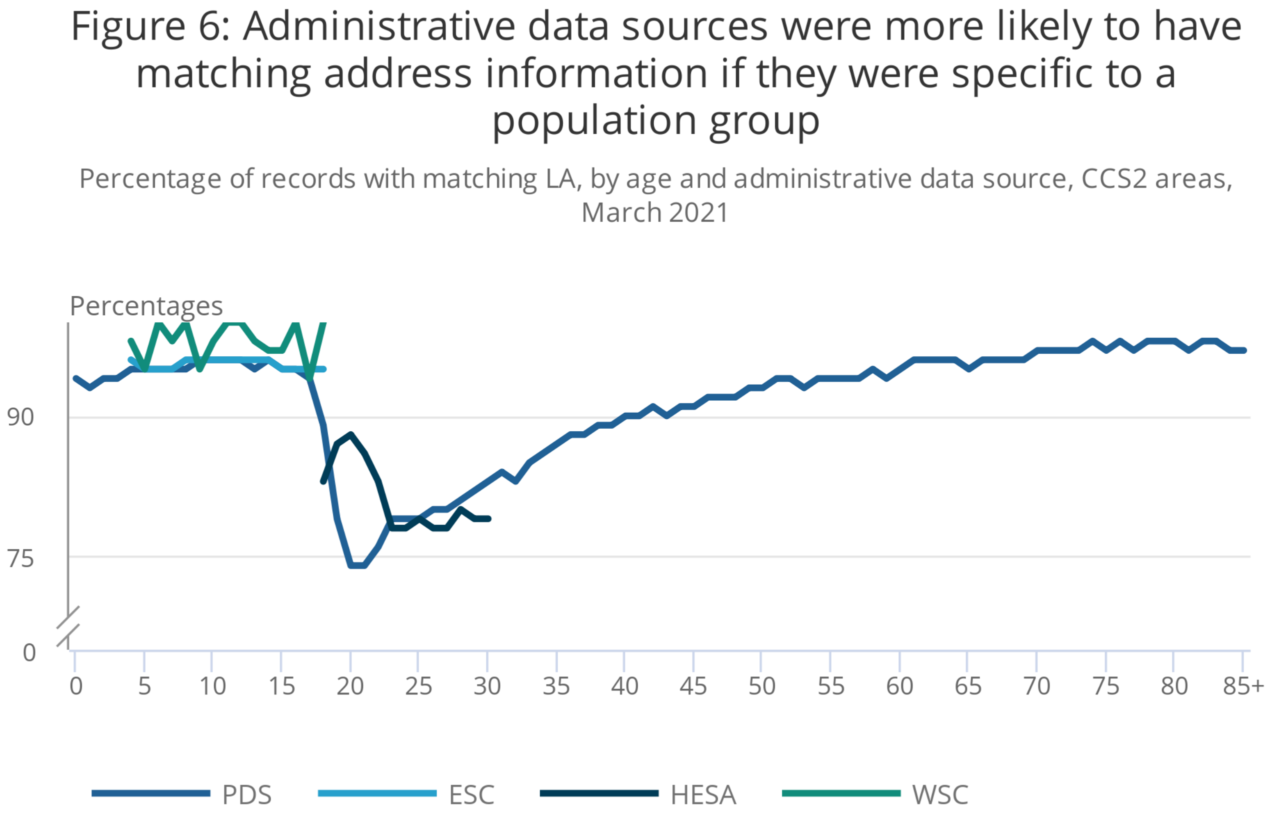
1. UPRN only available on PDS and ESC.

5 . Administrative data sources within the Demographic Index (DI) that are most likely to have matching address information, when compared with the census or CCS

We compared the Demographic Index (DI) address local authorities (LA) to with the Census-CCS (CC) address LAs and calculated the percentage of records from each administrative data source with a matching LA. This helps us understand which administrative data source is best to use when assigning address information, and how this varies by a record's age and sex.

Figure 6: Administrative data sources were more likely to have matching address information if they were specific to a population group

Percentage of records with matching LA, by age and administrative data source, CCS2 areas, March 2021



Source: Office for National Statistics

Notes:

1. HESA home LA used if HESA term time LA was not available.

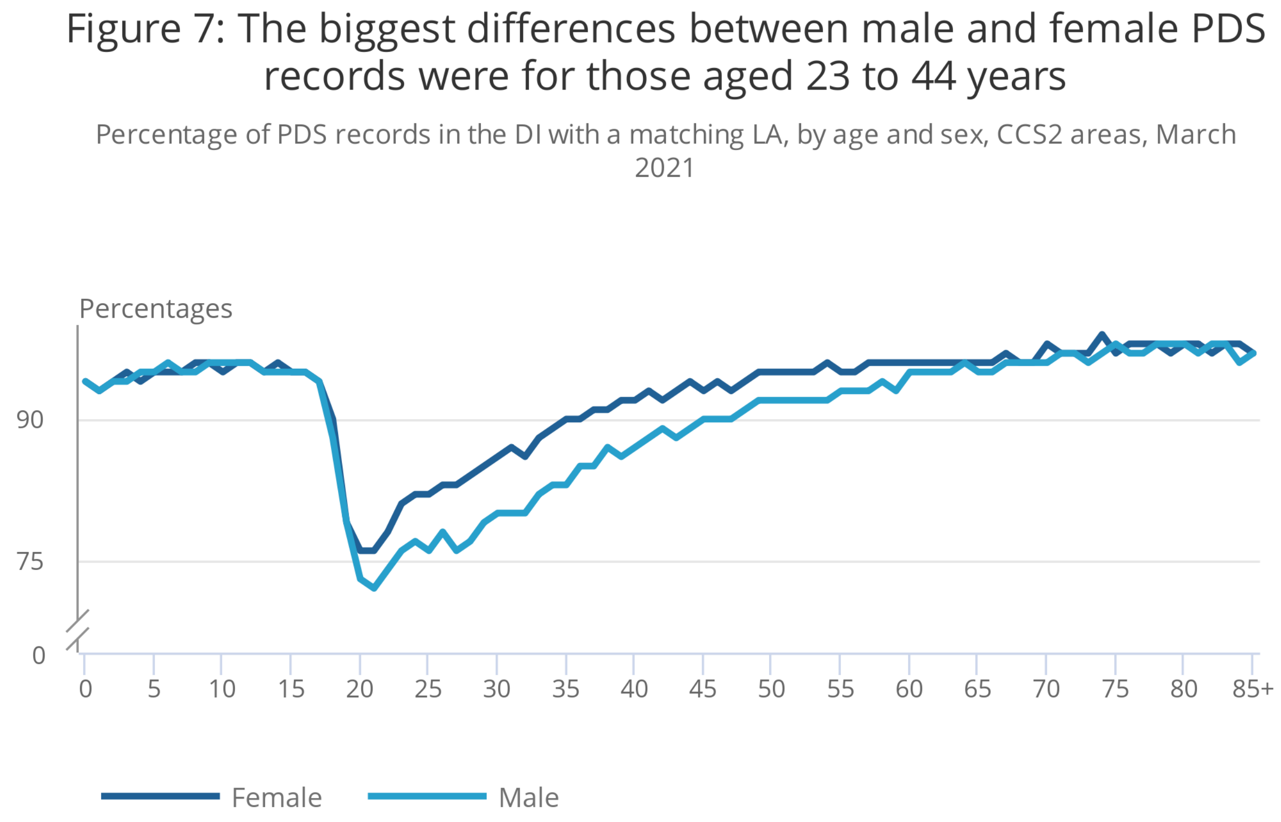
For school-aged children, a high percentage of English School Census (ESC), Welsh School Census (WSC), and Patient Demographic Service (PDS) records had a matching LA. Figure 6 shows that for those aged 19 to 22 years, many of whom are students, Higher Education Statistics Agency (HESA) had a higher percentage of records with matching LA than PDS. This suggests that administrative data sources are more likely to have matching address information when they are targeted to a specific population group.

For those of working-age, the percentage of PDS records with matching LA increased with age. Figure 6 shows that children and older people, who interact with NHS services more regularly, were also more likely to have a matching LA on the PDS. This suggests that administrative data sources are more likely to have matching address information for a specific population group, when that population group frequently interacts with the institution or service provider that collects the data.

We also looked at how the percentage of records with matching LA differed by sex and found no difference in ESC and WSC records. There was a small difference between male and female HESA records, with females more likely to have a matching LA. For those aged 20 to 60 years, there was a higher percentage of female PDS records with a matching LA. Figure 7 shows that the biggest differences between male and female PDS records were for those aged 23 to 44 years. However, for other age groups, there was little to no difference. This suggests that working-aged men may interact less frequently with NHS services, or move address more frequently, than other groups.

Figure 7: The biggest differences between male and female PDS records were for those aged 23 to 44 years

Percentage of PDS records in the DI with a matching LA, by age and sex, CCS2 areas, March 2021



Source: Office for National Statistics

6 . How well the Census-CCS (CC) and Demographic Index (DI) captures communal establishments in CCS2 areas

A communal establishment (CE) is a place providing managed residential accommodation, such as halls of residence, care homes, and prisons. Just over 1 million people live in a CE, according to our Communal establishment residents, England and Wales: Census 2021 bulletin. They are often a difficult group to capture (including in the census) because they often have large localised populations and are highly mobile.

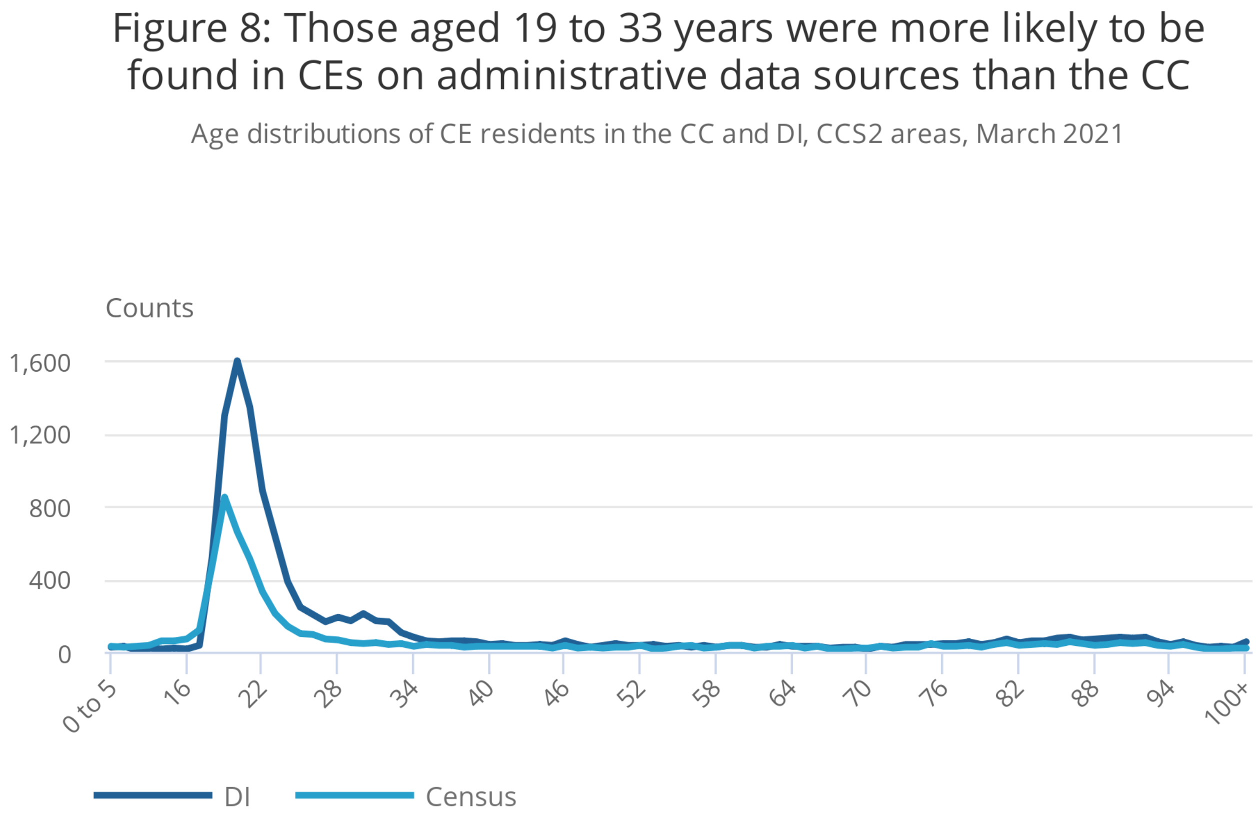
Our work will help inform future rule development for the Statistical Population Dataset (SPD) and provide insights for the dynamic population model (DPM). Note that our analysis was in CCS2 areas, which are skewed towards small CEs because of the Census Coverage Survey (CCS) sampling methodology.

The sample included 14,135 records. The aggregate age distribution showed that the DI counts were higher than CC for those aged 19 to 33 years, with the highest counts in the group aged 19 to 21 years. We think that using both PDS and HESA to include those in halls of residence is causing the overcoverage in these age groups. This might be a result of people not updating their address with service providers when they move from one place to another.

We saw undercoverage for those aged 13 to 17 years, and slight overcoverage for older adults. We found that sex ratios for the whole CE population and by CE type looked broadly similar on the CC and DI.

Figure 8: Those aged 19 to 33 years were more likely to be found in CEs on administrative data sources than the CC

Age distributions of CE residents in the CC and DI, CCS2 areas, March 2021



Source: Office for National Statistics

Notes:

1. Counts may not sum to stated totals because of rounding.

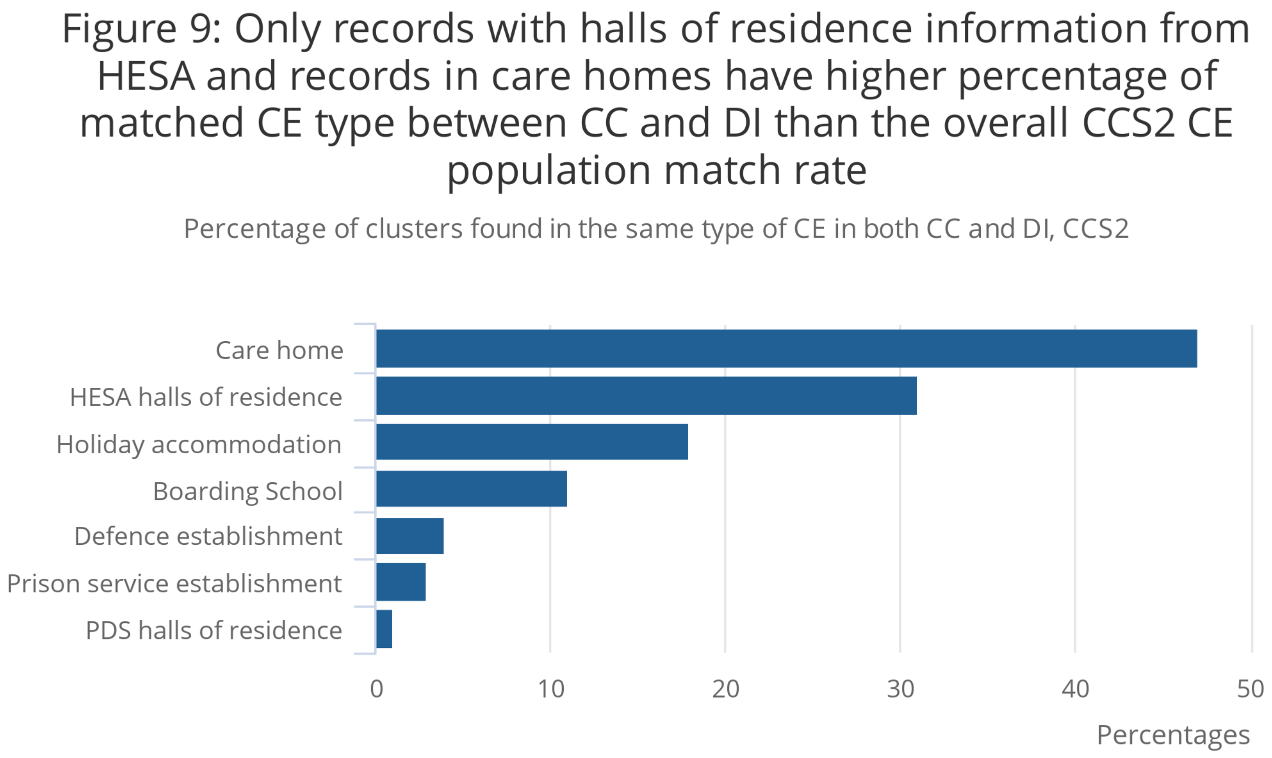
Record-level analysis of clusters assigned to a CE showed that:

- 25% of records were in a CE on both the DI and CC
- 19% of records were in a CE on CC only
- 56% of records were in a CE on the DI only

Breakdown by CE type showed that the match rate was unexpectedly low for all CE groups, and Patient Demographic Service (PDS) information on halls of residence scored lowest. Only care homes and Higher Education Statistics Agency (HESA) halls of residence information were above 25%. They were also the two largest CE groups and helped to bring the overall rate up.

Figure 9: Only records with halls of residence information from HESA and records in care homes have higher percentage of matched CE type between CC and DI than the overall CCS2 CE population match rate

Percentage of clusters found in the same type of CE in both CC and DI, CCS2



Source: Office for National Statistics

Notes:

1. There were many small counts which required suppression or grouping in line with disclosure control regulations.

The low match rates suggest that the selected administrative address information is often inaccurate, incorrectly assigning whether a person lives in a CE or a household. This could be because the address information is not updated to reflect moves in and out of CEs, or because we presently lack a sufficient way of identifying CEs in our administrative data's address information.

It is likely that the complications of collecting responses in CEs during the pandemic also influenced the match rates and the counts we saw on the CC. To learn about the effects of the pandemic on this, see our CE estimation and adjustment: Census 2021 methodology. The lockdowns caused abnormal movement of people and we saw lower numbers in some CEs, particularly education establishments. We plan to do further research to understand the results more fully for CEs and how these findings can be used to improve our SPDs and the input to the DPM.

As this analysis required records to be linked, we could only use actual census and CCS responses rather than using final adjusted census estimates. We believe that this may have disproportionately affected analysis conducted on CEs because of the variable response rates.

This has likely influenced the difference for those aged 18 to 33 years old in halls of residence on the two sources. However, our analysis also showed that PDS brought in large amounts of records where the residence type did not match the CC. Because PDS halls of residence were less consistent with census, we will review how we continue to use this data for our future methods.

7 . The Statistical Population Dataset (SPD)

The SPD aims to approximate the usually resident population, using administrative data in the Demographic Index (DI). As the DI includes data for records that would not be considered usual residents (for example, short-term migrants), the SPD applies rules to the DI to determine who it should include.

The linkage between the DI and Census-CCS (CC) enabled us to understand the SPD's coverage issues and the performance of its inclusion rules at a record level. By identifying those in the DI that matched to CC as a usual resident, we could determine if they should be included in the SPD. This enabled us to analyse three specific groups:

- records the SPD incorrectly excluded
- records the SPD incorrectly included
- records the SPD correctly excluded

The results of this analysis can be found in our Understanding quality of Statistical Population Dataset using the 2021 Census - Demographic Index linkage article.

8 . Glossary

Administrative data

Collections of data maintained for administrative reasons. Examples include 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.

Clerical review

Clerical review is a manual process whereby a person physically examines each combination of records and determines whether or not they match.

Data linkage

Data linkage is the process of joining together records that pertain to the same entity, such as a person or business. This can be automated or clerical. For more information, see our Developing standard tools for data linkage methodology.

Overcoverage

Overcoverage occurs when a record is counted more than once at the same location, more than once at a different location, counted in the wrong location, or is incorrectly included.

Residuals

Residuals are records that do not link between the two datasets, usually because they are only present on one or the other. However, some will be missed matches resulting from missing or incorrect information.

Undercoverage

Undercoverage occurs when a record is incorrectly excluded from data.

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.

9 . Data sources and quality

Census 2021

Response to Census 2021 was very high, at around 97%. The quality of the census estimates were considered very high after applying a coverage adjustment process that accounted for non-response and incorrect responses. This process is detailed in our Coverage estimation for Census 2021 in England and Wales methodology. However, since our analysis for Census Coverage Survey 2 (CCS2) required the matching of individual records, we could only include those who responded to Census 2021 or the Census Coverage Survey (CCS). In addition, CCS areas, and by extension the CCS2, typically did not include large communal establishments (CEs) as they are a subject to a separate estimation process. This information can be found in the UKSA's Estimating Populations in Large Communal Establishments (PDF, 208KB). The CCS sample is designed to target areas of low census response, as explained in the UKSA's 2021 Census coverage survey: sample allocation strategy (DOCX, 2MB). As a result, both Census 2021 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.

Administrative data

Administrative data is collected continuously throughout the year and a snapshot taken at specific times. For the data used in this publication, none of the extracts were taken on Census Day (21 March 2021), so there is a possibility of time-lag error. Where possible, we have addressed this. For example, we have removed babies born after Census Day from the analysis discussed in Section 7.

Where possible, we have used administrative data that covers 2021; 2020 data for HESA was used in the initial linkage, but 2021 data has since been used to provide more relevant data for Sections 4, 5, and 6.

Sections 4, 5, and 6 did not use DWP's Customer Information System (CIS) data, to ensure compliance with data sharing agreements held.

More on the administrative data sources used in this paper, and how they were used to support the census process, can be found in our Administrative data used in Census 2021, England and Wales .

10 . Future developments

Conducting this research has already provided us with a much better understanding of the Demographic Index (DI). We will continue to develop this understanding by expanding the analysis outlined in this article, including:

- using census address one year ago, to better understand how lags in address changes feeding through into the administrative data impacts our Statistical Population Dataset (SPD)
- further analysis to provide evidence to help address the need for a robust coverage adjustment method for the SPD (as set out in the UKSA's SPD Estimation Options (PDF, 1.3MB) to support its use in the dynamic population model (DPM)
- learning more about households in England and Wales and how well we can estimate household size and composition using administrative data
- learning more about communal establishment (CE) residents and specific population groups, such as students, non-state school pupils, and UK armed forces, and how we could improve our coverage of them within administrative data

The impact of this project will also have implications across the Office for National Statistics (ONS), outside of the population and social statistics transformation programme. The complexity of the linkage will allow us to understand more about the quality of existing linkages within ONS, as well as how clerical review can help us to achieve better results when linking these types of data.

11 . Related links

Transforming population statistics, comparing 2021 population estimates in England and Wales
Article | Released 28 February 2023
Evaluating progress towards a transformed population statistics system, using comparisons between census-based and admin-based population estimates and Census 2021.

Developing Statistical Population Datasets, England and Wales: 2021
Article | Released 28 February 2023
Aggregate comparisons between the Statistical Population Dataset version 4.0 (v4.0) and Census 2021.

Understanding quality of Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage
Article | Released 28 February 2023
Analysis of Statistical Population Dataset version 4.0 2021 using a linkage between Census 2021 and the Demographic Index.

Admin-based population estimates: provisional estimates for local authorities in England and Wales, 2011 to 2022
Article | Released 28 February 2023
Admin-based population estimates for all local authorities in England and Wales from the dynamic population model.

12 . Cite this article

Office for National Statistics (ONS), released 1 March 2023, ONS website, article, Understanding quality of linked administrative data sources in England and Wales, using the 2021 Census - Demographic Index linkage