Linkage methods for Census 2021 in England and Wales

Matching Census 2021 to the Census Coverage Survey (CCS) using deterministic, probabilistic, associative, machine learning and clerical methods.

Contact:
Census customer services
Census.customerservices@ons.gov.uk
+44 1329 444972

Release date:
15 December 2022

Next release:
To be announced

Table of contents

1 . Main points

- Census 2021 matching exceeded quality targets, as detailed in Section 7: Results and quality assurance.
- We achieved matching precisions of 99.96% and 100.00% for person and household records respectively, exceeding our target of at least 99.90% for both.
- We achieved matching recalls of 99.96% and 99.78% for person and household records respectively, exceeding our target of at least 99.75% for both.
- Matching was completed in a four-week period from receipt of the final census and Census Coverage Survey (CCS) results.
- We achieved automatic match rates of 93.1% and 95.6% for person and household records respectively.
- We achieved final match rates of 97.3% and 97.1% for person and household records respectively.
- Matching methods included a combination of deterministic, probabilistic, associative, machine learning and clerical; these were all improved iteratively throughout the tuning stage.
- It would not have been possible to achieve both precision and recall targets without the use of clerical matching.
- We developed the clerical matching system (CMS) and a team of 50 clerical matchers used this successfully throughout all stages of matching.

2 . Overview

To measure the coverage in Census 2021 in England and Wales, the Office for National Statistics (ONS) conducted a Census Coverage Survey (CCS) across England and Wales. This was a voluntary survey carried out independently of the census, designed to measure coverage in the census.

The CCS was used to estimate the population counted and missed by the census and adjust the census database for those estimated to have been missed (Whitworth and others, 2018). This is part of the Design for Census 2021.

Matching Census 2021 to the CCS is a vital part of the coverage assessment and adjustment processes. These processes rely on records from the CCS being correctly paired with census records that correspond to the same person (the same goes for matching households).

The fundamental problem in matching is that a single person or household may have non-identical records in the census and the CCS. This can occur when there is missing or incomplete information in one of the records, because of spelling mistakes, scanning errors, changes to names or addresses, or other mistakes. A description of the CCS and how it is used to produce quality population estimates can be found in Maximising the quality of Census 2021 population estimates.

The quality of the matching is measured by precision and recall, which are defined as follows:

$$precision = \frac{\text{number of true matches}}{\text{number of true matches} + \text{number of false matches}}$$

Precision is a measure of the accuracy of the matches that have been made:

$$\text{recall} = \frac{\text{number of true matches}}{\text{number of true matches} + \text{number of missed matches}}$$

Recall is a measure of the proportion of matches that have been made out of all the possible matches. For this work, precision and recall targets were 99.90% and 99.75% for both person- and household-level matching. These precision and recall targets are extremely high compared with typical linkage quality targets, as any linkage error directly affects the quality of the census estimates. The final matched and unmatched records are used for the process of adjustment using dual system estimation (DSE). More information on DSE can be found in Trout, Catfish and Roach (PDF, 818KB).

3 . Pre-linkage

All linkage projects have a number of stages and the first is to clean and prepare the data to maximise the linkage accuracy. Once collected, census and Census Coverage Survey (CCS) data records were passed through a number of validation and cleaning processes. More information on these processes can be found in Maximising the quality of the 2021 population estimates – cleaning the data.

Before matching the census and CCS, extra pre-processing tasks were carried out to maximise the quality of the data linkage. These tasks included cleaning matching variables (for example, name, date of birth, address), deriving new matching variables (for example, nickname, postcode area, house number, year of birth) and removing out of scope records, including:

- duplicates missed by the resolve multiple responses process
- CCS records not within a CCS collection area
- individuals born after Census Day
- records missing all names and date of birth
- students at non-term time addresses

4 . Methods used for census to CCS matching

The following methods listed were used to match Census 2021 to the Census Coverage Survey (CCS). Further information on these processes is provided later in this section and in this Methodological Assurance Review Panel (MARP) paper: Quality Control and Quality Assurance Strategy for 2021 Census to CCS Person and Household Matching (PDF, 1.92MB) (Annex C).

Deterministic matching was used to automatically match persons and households that satisfied a series of rules called matchkeys. The matchkeys capture exact matches, but also allow for error in some of the variables.

Probabilistic matching was also used to match person records. Pairs of census and CCS records that were awarded a probabilistic score above a threshold were accepted automatically if they were also matched deterministically and were otherwise sent for clerical resolution. Pairs awarded a probabilistic score below the threshold were rejected automatically.

Pairwise clerical matching was used to determine the true match status for pairs of census and CCS records, which were not matched automatically. It was also used to resolve clusters of matches where more than one CCS record matched to the same census record, or the other way around. This clerical matching was carried out using a custom-built clerical matching system (CMS).

Associative clerical matching was used to present matched households containing unmatched persons in the CMS. This enabled optimum within-household person linkage. Unmatched households containing matched persons were also presented in the CMS to determine whether the households should also be matched.

Presearch clerical matching was used to present the most likely census candidates for all remaining unmatched CCS persons from the previous stages. Presearch used a combination of probabilistic linkage and a machine learning gradient boosted decision tree model. Person matches could then be made in the CMS, and CCS persons still unmatched after this point were declared as unmatchable records.

Clerical searchwas used after linkage to estimate the false negative rate (person matching only). Clerical matchers were presented with an unmatched CCS record and a search screen that uses exact and wildcard searching to seek a matching record in the census dataset.

Deterministic matching

Deterministic matching was used to automatically match as many CCS records to the census as possible. A series of rules called matchkeys were initially developed using 2011 Census and CCS data and then updated once 2021 data were made available.

A matchkey contains a set of criteria that two records must meet if they are to be declared a match. If the two records disagree on any of the criteria, then the pair is classified as a non-match. For example, the third deterministic matchkey required forename, surname, date of birth, sex and postcode to be the same in the census and CCS record.

Deterministic algorithms often use hierarchical matchkeys, which means that the matchkeys are ordered by strictness, and once a census or CCS record is matched it will not then be considered for matching in the remaining matchkeys. Because of the high precision and recall targets for this linkage, hierarchical matching was not used, and all records were passed through all matchkeys. This ensured matches were not missed just because they were made on a weaker matchkey. It also meant there was the possibility of making conflicting or non-unique matches within or across matchkeys, as shown in Table 1.

Table 1: Example of outputs following the Census to Census Coverage Survey deterministic matching

Census Record	CCS Record	Matchkey	Unique Match	Decision
A	B	1	No	Clerical resolution
A	C	15	No	Clerical resolution
D	E	6	Yes	Auto-match

Source: Office for National Statistics – Census Coverage Survey

Table 1 shows that census record A has been matched to CCS records B and C. This suggests that B and C could be the same person. These records are sent to the CMS to confirm that the matches are correct, and B and C are in fact duplicates. Census record D has been uniquely matched to CCS record E, therefore these records can be automatically matched without clerical input.

In the final run of the deterministic matching, 35 matchkeys were used to match person records. These primarily used variables derived from name, date of birth, sex and address, with different amounts of error allowed within each matchkey. Three of the 35 person matchkeys were not used to automatically match records, as their false positive rates were deemed too high. Instead of removing these matchkeys (and potentially missing these matches in later stages), all matches made by these matchkeys were clerically reviewed in the CMS.

Deterministic matching was also used for matching households. A household match was defined to be the same space with at least one person in common. Twenty-nine matchkeys were developed, using a combination of household (accommodation type, tenure and so on) and person information. Our GitHub page (ccslink) contains the final person matchkey list and household matchkey list.

Probabilistic matching

A Fellegi-Sunter probabilistic linkage algorithm (Fellegi and Sunter, 1969) was applied to all census and CCS person records, regardless of their match status from the deterministic stage. This meant that the results from the two stages could be compared against each other and used to iteratively improve both methods.

Firstly, similar-looking records were brought together using a process called blocking. Blocking uses a series of rules similar to deterministic matchkeys, however, these are much looser and bring together a mixture of non-matches and true matches. In 2011, over 98% of all census-CCS person matches agreed on the enumeration postcode, therefore our main blocking pass used this variable only.

Five other blocking passes were applied, using a combination of enumeration postcode, alternative postcode and Census Day postcode (for CCS responders who have moved since Census Day). True matches where the postcodes do not agree could be found in later stages.

Every candidate pair produced from the blocking was run through the probabilistic linkage algorithm. The algorithm used agreement and disagreement weights for each matching variable to generate a final match score for all pairs. To calculate these weights, two probabilities must be estimated for each matching variable. These are defined as:

- *M* probability: the probability that a variable agrees on the census and the CCS, given the pair are a true match probability: the probability that a variable agrees on the census and the CCS, given the pair are a true match
- *U* probability: the probability that a variable agrees on the census and the CCS, given the pair are not a true match

M values were initially estimated using all matches made deterministically. However, as we started to make more matches clerically, active learning was used to update the M values. U values are calculated using a random sample of census – CCS pairs; therefore, they did not need to be updated when new matches were made.

The matching variables chosen for this linkage were forename, surname, date of birth and sex. Clerical resolution was used to find a threshold that split potential matches from non-matches.

Once scores for all blocked census – CCS pairs were calculated, the results were combined with the deterministic matching results and the following steps were applied:

1. record pairs automatically matched in the deterministic stage remained automatically matched unless an alternative match for the same records scored above the probabilistic threshold; in this case, both pairs were sent for clerical resolution
2. all probabilistic matches that scored above the threshold and were not matched deterministically were sent for clerical resolution
3. all probabilistic matches that scored below the threshold and were not matched deterministically were discarded

New matches from the probabilistic algorithm (step 2) were always sent for clerical resolution as a small number of incorrect matches existed in even the highest scoring candidates. Once all clerical matches were resolved, the accepted matches were combined with the automatic matches.

The deterministic and probabilistic linkage stages were run multiple times during the live matching period, as not all census and CCS data were available from the very start. Each time we received a new cut of data containing extra records, the pipeline was re-run and new matches could be made either automatically or clerically.

Associative matching

Associative matching combined person and household decisions from the deterministic and probabilistic stages to identify new possible matches that could be made. Firstly, any matched households that contained one or more unmatched person records were sent for clerical resolution in the CMS. The matcher was presented with the household match, any person matches already made, and all remaining unmatched person records from both households. New person matches could be made, or existing person and household matches could be broken.

Secondly, household pairs (plus their person records) were sent to the CMS for clerical resolution if they were unmatched but contained one or more person match. The matcher could then decide if the households were a match and make or break person matches.

Associative matching allowed clerical matchers to easily match person records that would be challenging to match using their name, date of birth, sex and geography alone. It also allowed matchers to match households with errors in postcodes and addresses, usually caused by scanning errors (paper responses), field collection errors or address matching errors.

Presearch matching

All unmatched CCS person records from the previous stages were run through presearch to find the most difficult matches still to be made. Presearch is a score-based method that uses a combination of a probabilistic linkage model and a gradient-boosted decision tree (GBT) model.

The probabilistic part of presearch was adapted from the previous, simpler probabilistic stage. To capture the harder matches, extra matching variables were included and more comparisons between different variables were made (for example, comparing surname with middle name). Postcode area was used as a single blocking variable to reduce the search space for finding matches, although an extra blocking pass with no geography was also used.

The GBT supervised machine learning method combined predictions from several different decisions trees into a single model. This single model is an improvement on the previous models as it minimises the overall prediction error. The probabilistic and GBT models were both developed using the 2011 Census and CCS data and then improved during the 2021 tuning period to account for changes in data quality.

To combine the results of the two methods, the highest scoring census candidates from each method were selected (for example, select highest scoring probabilistic record, then select highest GBT record and so on) until 15 unique candidates were reached. We found that taking the combined top 15 census records performed better than taking the top 15 from just one of the methods, as more true matches were found.

The top 15 census candidates were presented alongside each unmatched CCS record in the CMS. Clerical matchers could then match the CCS record to one (or more) of the census records or declare it as unmatchable, meaning that we believe the CCS person did not respond to the census.

5 . Clerical matching

Clerical matching system

A team of 50 clerical matchers, which included nine expert matchers and four supervisors, were employed to complete the census to Census Coverage Survey (CCS) clerical matching tasks described in this article. Every stage of matching resulted in some records requiring clerical resolution in the clerical matching system (CMS). Clerical matching could be carried out in one of four journeys:

- individual journey – used for matching pairs or clusters of person records
- household journey – used for associative matching
- presearch journey – used for presearch matching
- search journey – used for quality assurance tasks

In all these journeys, the clerical matcher can choose to view additional information, household view, or a PDF of the paper form (if it exists). Additional information comprises further variables such as alternative addresses, marital status, ethnicity and occupation, which are not shown on the first screen. Household view shows the other people in the household and the relationships between them. The clerical matcher can also choose to send the records to an expert matcher for a second opinion, or to report an issue.

Many matching decisions in the CMS were subjective. This was because of a variety of factors including missing information, common names and so on. It was found that higher quality clerical matching was achieved by sending all pairs of records to the CMS twice, and then sending any pairs with conflicting decisions (for example, matched first time around, not matched second time around) back for a third opinion. In this way, each clerical decision was a "best of three".

In addition to the best of three rule, approximately 15% of all matches made in the CMS were reviewed again by an expert matcher. The expert matcher had the ability to change the final decision if they thought the clerical matcher had made a mistake. Additional training was given if this was the case.

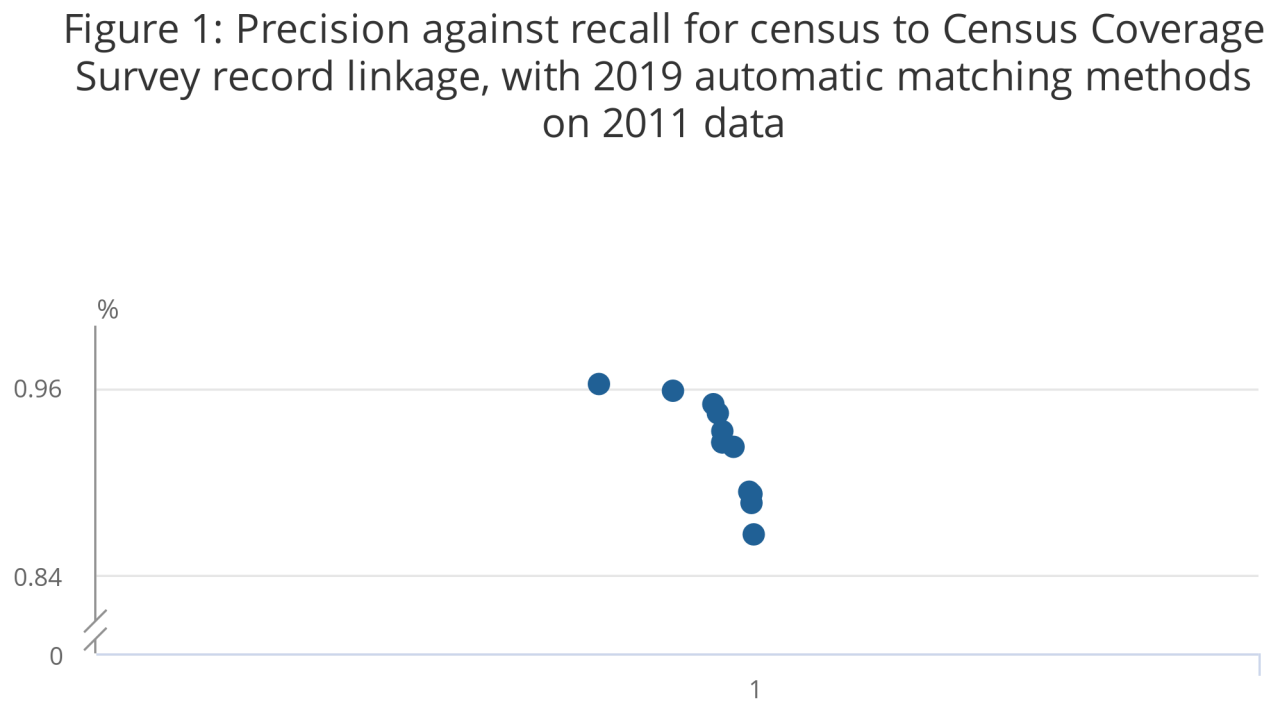
Why clerical matching was needed

In the 2011 Census, the matching of records between the census and CCS required significant clerical effort to find matches that could not be found by automatic means. Since then, we have continued to develop methods of automated matching, including machine learning approaches.

These automatic matching improvements have resulted in increases in both precision and recall. For example, in 2011 a recall of 70% was achieved using automated methods for matching person records, and our improvements have now enabled us to achieve a recall of 90% using the same data, with no loss of precision. While this is good, the recall is still significantly short of the 99.75% target.

The Office for National Statistics (ONS) collaborated with the Alan Turing Institute in 2019 to investigate alternative methods that may enable a fully automated process for Census 2021 to CCS matching. A number of alternatives were considered (for example, decision trees, expectation maximisation without the conditional independence assumption and so on). However, it was concluded that the combination of deterministic matching and Fellegi-Sunter probabilistic matching remained the best approach for matching these particular data, and resources should be concentrated on continuing to improve the existing methods already developed.

Figure 1: Precision against recall for census to Census Coverage Survey record linkage, with 2019 automatic matching methods on 2011 data



Source: Office for National Statistics – Census Coverage Survey

The ONS and the Alan Turing Institute carried out precision and recall analysis as part of this work. Figure 1 shows precision against recall for our improved Census to CCS automatic matching methods on 2011 data. Data points show different thresholds for considering record pairs a match based on Fellegi-Sunter scoring.

Figure 1 suggests that it is not possible to achieve 99.90% precision and 99.75% recall using our combined deterministic and probabilistic approach. Relaxing the probabilistic threshold does find more true matches to increase recall, however, this results in the precision requirement no longer being met.

6 . Census to census matching

In addition to linking Census 2021 to the Census Coverage Survey (CCS), we linked samples of the census data to the census in order to enable estimation of the number of duplicate responses in the census.

An inverse sampling technique was used to ensure that the samples of census data that we matched were large enough to enable estimation of duplication rates for particular demographics of interest with a coefficient of variation of less than 10%. Similar to the census to CCS matching we employed several matching techniques including probabilistic, deterministic and clerical to ensure that the accuracy of the linkage met the high accuracy requirements. Full details of the methods employed for census to census matching can be found in this Methodological Assurance Review Panel (MARP) paper: Census to census matching strategy 2021 (PDF, 386KB).

7 . Results and quality assurance

Quality assurance tasks were carried out during the live matching period to ensure that all targets were being met. These tasks were mostly carried out in the clerical matching system (CMS) and included:

- quality assurance of deterministic matchkeys
- quality assurance for finding probabilistic threshold
- quality assurance for tuning of the presearch algorithm
- false positive searching – checking for false positives (false matches) in a sample of clerically matched records
- false negative searching – checking for false negatives (missed matches) in a sample of unmatched Census Coverage Survey (CCS) records

Once all matching was complete, quality metrics precision and recall were estimated using the matched and unmatched records. To calculate the precision of the census to CCS matching, samples of both automatic and clerical matches were clerically reviewed to check for false matches. To calculate recall, a sample of unmatched CCS records was taken, and a census match was searched for in the CMS. Calculations for these metrics (person matching) are presented in Tables 2 and 3.

Table 2: Precision results for 2021 Census to Census Coverage Survey person matching
Precision calculated for automatic, clerical and all matches

Matching method	Total matches made	Total matches sampled	False positives in sample	Total false positives estimated	Precision	Confidence Interval (95%)
Automatic	425,859	20,000	1	21	0.99995	[0.99968, 1.00000]
Clerical	22,527	2,000	13	146	0.99352	[0.98874, 0.99629]
Overall	448,386	-	-	167	0.99963	[0.99913, 0.99981]

Source: Office for National Statistics – Census Coverage Survey

Table 3: Recall estimated for 2021 Census to Census Coverage Survey person matching

Total unmatched CCS records	Total unmatched CCS records sampled	Total matches found in sample	Total false negatives estimated	Total true positives	Recall	Confidence Interval (95%)
12,221	800	12	184	448,219	0.99959	[0.99928, 0.99977]

Source: Office for National Statistics – Census Coverage Survey

As expected, false positives were made more frequently during clerical matching compared with automatic matching. Of all clerical matches, 0.648% were estimated to be false positives, whereas only 0.005% of automatic matches were estimated to be false positives. Table 3 shows that only 12 true matches were found in a sample of 800 unmatched CCS records, resulting in an overall recall of 99.96%. Table 4 shows that the overall accuracy of Census 2021 to CCS matching exceeded the targets set for precision and recall for both person and household matching.

Table 4: Final precision and recall estimated for 2021 Census to Census Coverage Survey person and household matching

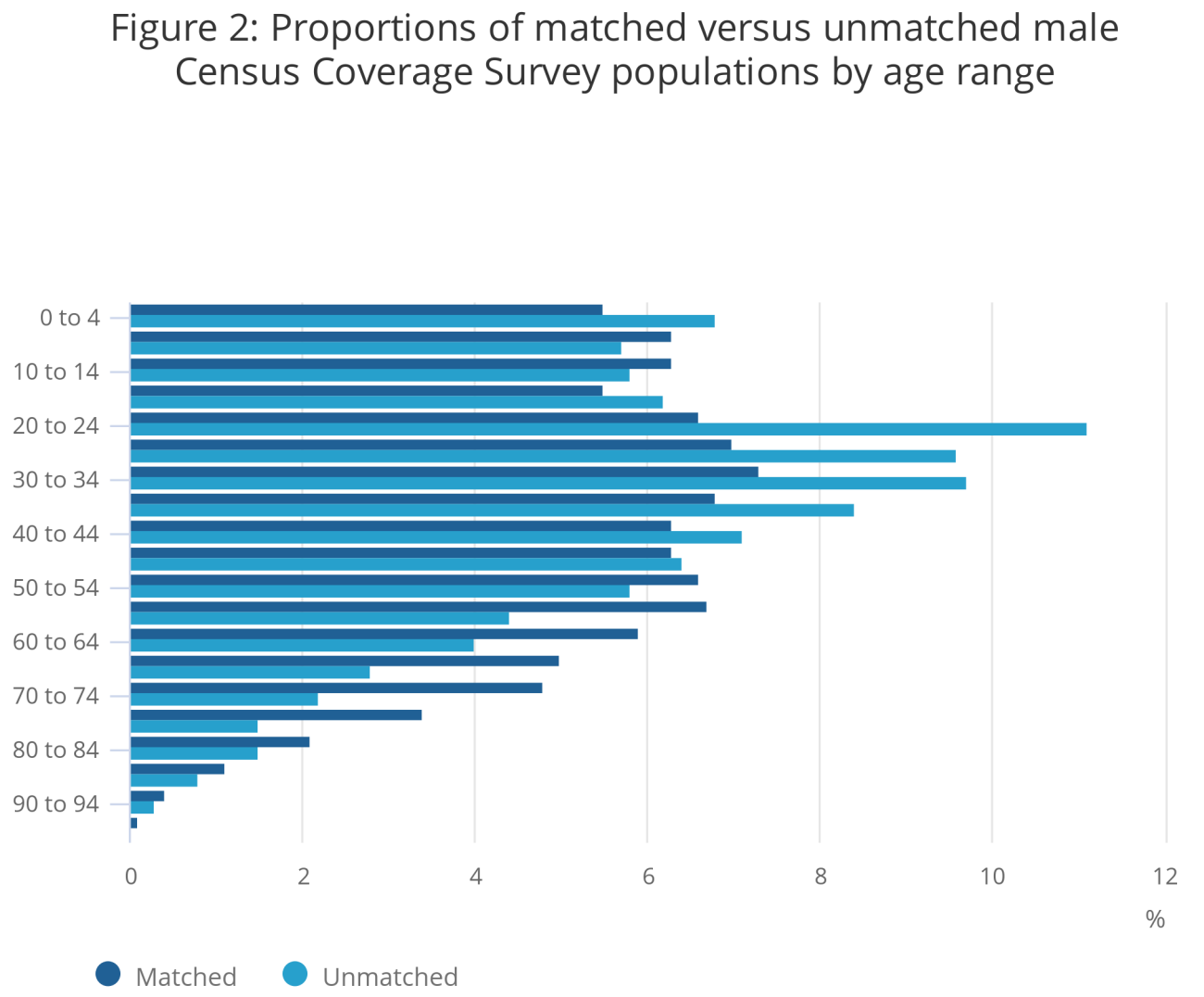
	Precision (Target)	Recall (Target)
Persons	99.96% (99.90%)	99.96% (99.75%)
Households	100% (99.90%)	99.78% (99.75%)

Source: Office for National Statistics – Census Coverage Survey

False positives and false negatives that occur during the linkage process can lead to biased results. Insight on the potential bias of a linkage may be drawn by exploring the characteristics of the matched and unmatched records. It is difficult to unpick linkage bias from non-response bias; however, our very low number of false negatives means that we expect most bias to be non-response bias.

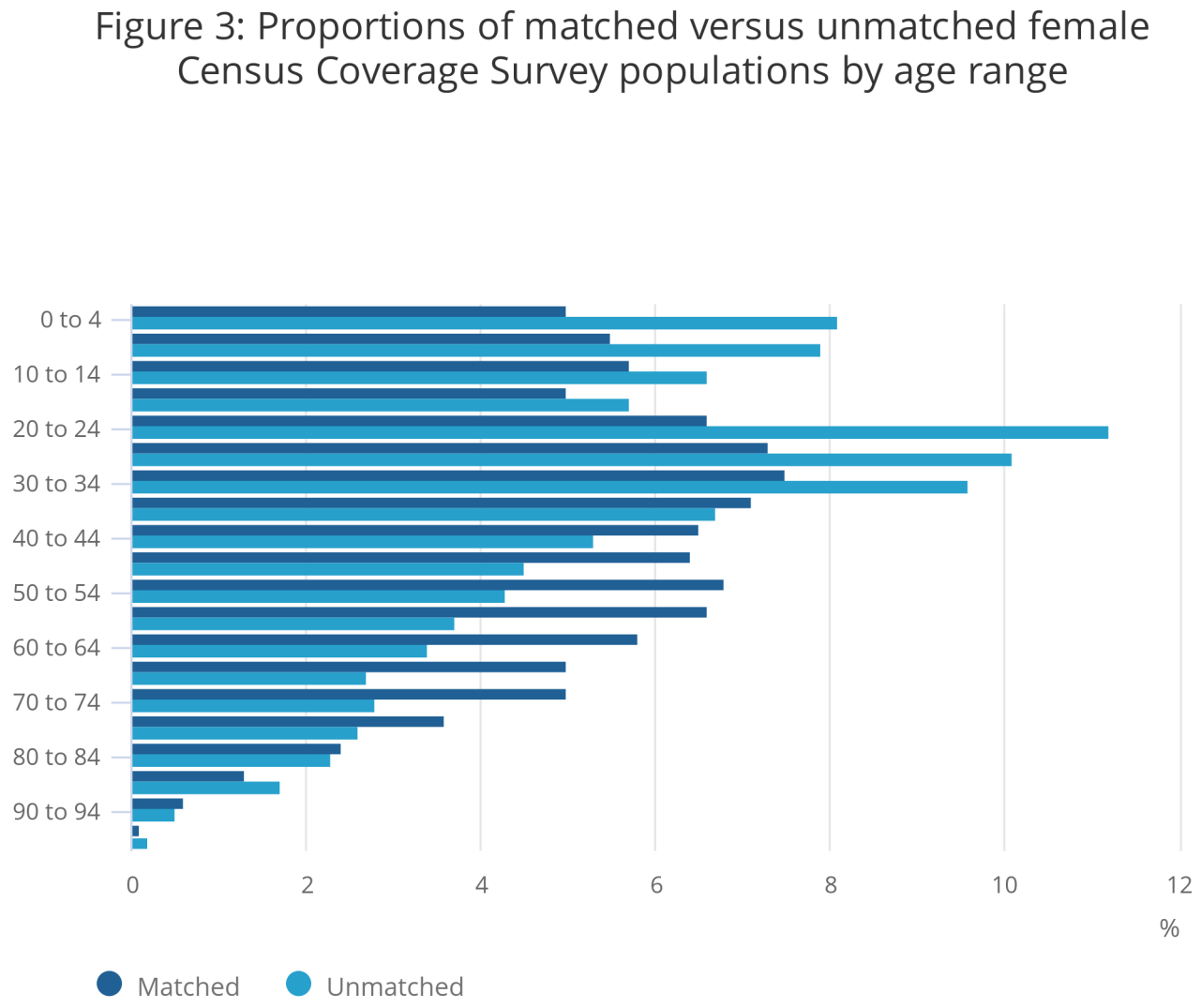
Figures 2 and 3 show the age distributions of the matched CCS data versus the unmatched CCS data for males (Figure 2) and females (Figure 3). There is more than a 3% increase in the proportions of 20- to 24-year-olds in the unmatched data compared with the matched data for both males and females. Likewise, both figures show a decrease in proportions of 45- to 84-year-olds in the unmatched data when compared with the matched data. We know from past censuses (2011 General report for England and Wales chapter 8 (PDF, 351KB)) that 20- to 24-year-olds are less likely to respond to the census compared with other age groups, therefore this non-response bias was expected.

Figure 2: Proportions of matched versus unmatched male Census Coverage Survey populations by age range



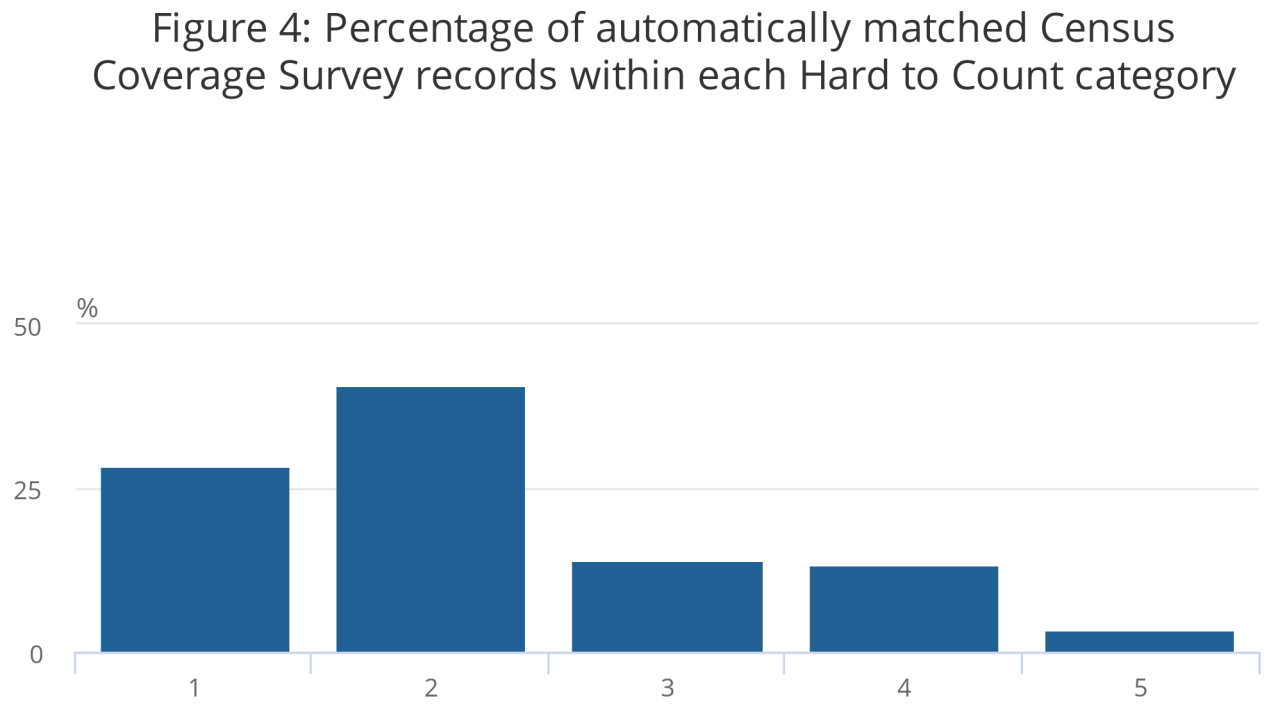
Source: Office for National Statistics – Census Coverage Survey

Figure 3: Proportions of matched versus unmatched female Census Coverage Survey populations by age range



Source: Office for National Statistics – Census Coverage Survey

Figure 4: Percentage of automatically matched Census Coverage Survey records within each Hard to Count category

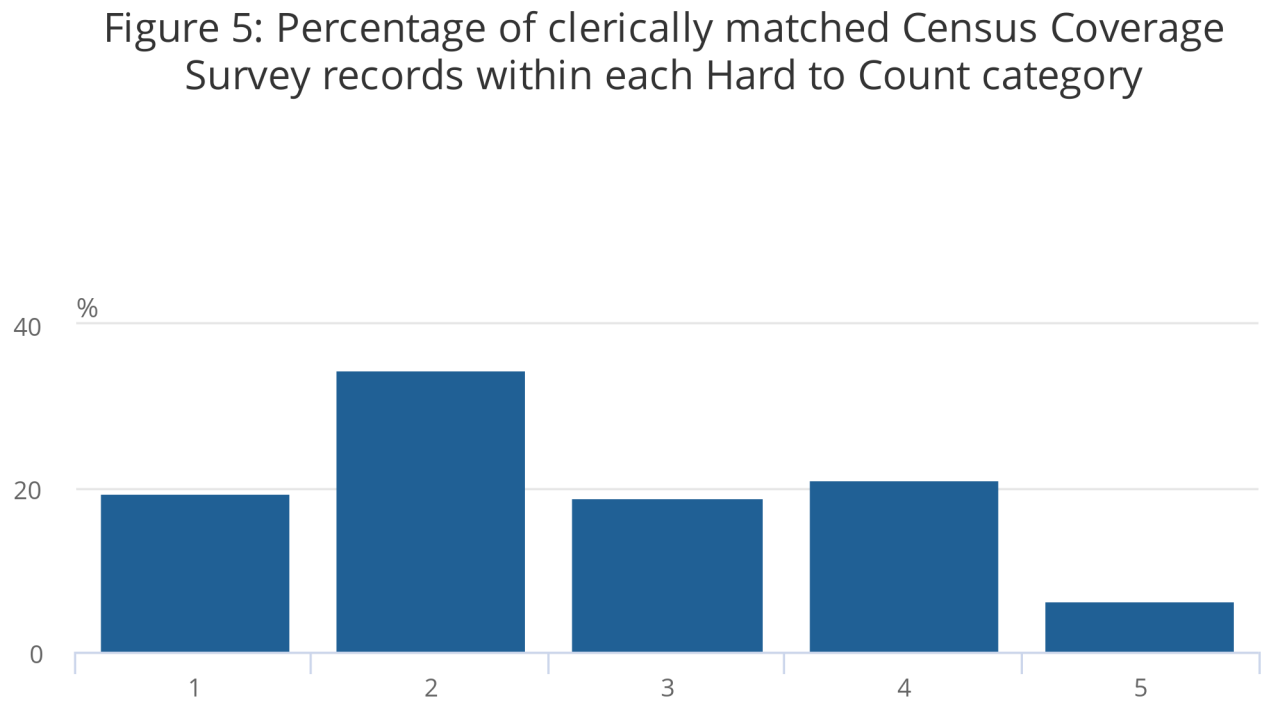


Source: Office for National Statistics – Census Coverage Survey

We can also compare the distributions of the automatic and clerical matches by different characteristics. Figures 4 and 5 show the percentage of CCS person records that sit in each hard-to-count (DOCX, 1,273 KB) (HTC) group for automatic matches (Figure 4) and clerical matches (Figure 5). The HTC index rates postcodes on a scale from 1 to 5, with 5 being the areas that are the hardest to enumerate.

The HTC distribution of clerical matches is clearly different compared with the distribution of the automatic matches, as the clerical matches contain a higher proportion of matches from typically harder to match areas (HTC 3, 4 and 5). Without these clerical matches, we would not have achieved our recall target, and we would have had substantial linkage bias as the HTC distribution of the false negatives would have been significantly different to that of the true positives. Similar trends can be seen when looking at other variables (for example, ethnicity).

Figure 5: Percentage of clerically matched Census Coverage Survey records within each Hard to Count category



Source: Office for National Statistics – Census Coverage Survey

In addition to the quality assurance methods mentioned so far, census to CCS person matching quality was also evaluated using the Office for National Statistics (ONS) Longitudinal Study (LS).

Firstly, both Census 2021 and CCS datasets were filtered to only include individuals residing in a CCS area and with an LS birth date. The datasets were then both matched to the LS dataset – this could be done via NHS number, as both the census and CCS had previously been independently matched to the Personal Demographic Service (PDS) dataset. The census and CCS were then matched on LS number to create a LS proxy census to CCS matched dataset. This matched set could be compared with the original census to CCS matched set to identify possible missed matches in the original matching. More information on the evaluation can be found in Annex B of Quality Assurance Strategy for 2021 Census to CCS Person and Household Matching (PDF, 1.92 MB).

The LS proxy matched set contained 5,528 matches, of which only three were not captured in the original census to CCS matched set. From this we could estimate a recall of 99.95%, reaffirming that our recall quality target has been met.

8 . Related links

Maximising the quality of Census 2021 population estimates
Methodology | Last revised 9 December 2022
How we maximised the quality of Census 2021 population estimates during the processing and quality assurance of the final statistics.

9 . Cite this methodology

Office for National Statistics (ONS), released 15 December 2022, ONS website, methodology, Linkage methods for Census 2021 in England and Wales.