Scotland’s Census 2022

Census–CCS Person Linking

June 2020

Contents

1. Plain English Abstract

All households in Scotland are required to complete a census return for all usually resident persons. However, sometimes people are missed. In order to avoid underestimating the population as a result of this, a sample of areas are surveyed again in a Census Coverage Survey (CCS). By comparing the responses from the CCS to those from the census, we can estimate how many people are missing from the census.

In order to compare the census and CCS we need to identify which people respond to both. We therefore link the people on the CCS to those on the census. To check that records from the census and CCS represent the same person, we compare the name, date of birth, sex and address that is recorded on the two questionnaires. We then manually check all the CCS records that have not linked to the census, against similar census records, to ensure we do not miss any matches.

2. Abstract

All households in Scotland are required to complete a census return for all usually resident persons, although sometimes people are missed. In order to avoid underestimating the population as a result of this, a sample of areas are surveyed again in a Census Coverage Survey (CCS). This is carried out a few weeks after census day, independently of the census. The records from the CCS are linked to those from the census in order to count the number of people appearing on both, and the number appearing only on the census, or only the CCS. Dual-system estimation (DSE) uses these counts to estimate the total population.

This process relies on an accurate count of the people appearing on both sources. To find this each CCS record is compared to census records in terms of the name, date of birth, postcode and sex. Comparing every CCS record to every census record would be unfeasible. Therefore we separate the data into blocks[1] and compare the CCS records to the census records that are in the same block. For

[1] A block is a set of records that are only compared to other records in the same block during linking. See Steorts et al. (2014) for a discussion of blocking.

example, records in one postcode might be compared only to other records in that postcode. To avoid missing matches, linking is repeated using a range of blocks such as postcode sector, date of birth, and parts of the name.

Each comparison is scored according to the similarity of the different information (first name, last name and date of birth). These scores are used to categorise the link and assign a distance score. These distance scores are combined with a score derived from the similarity of the postcodes to give an overall distance score. If a CCS record links to just one census record with a better overall distance score than to any other census record then that link is automatically made. Remaining CCS records are all then manually compared to the list of census records they are linked to (ranked by the overall strength of the link) to find any remaining matches. In this way we can reduce both the false positives (links made that are not matches) and false negatives (matches that have not been linked). The method was tested using the data from 2011 and also the 2019 rehearsal. The testing suggested that the net systematic error in the population estimate from false positives and false negatives would be less than 0.1 per cent (well within the 0.5 per cent Key Performance Indicator for bias[2]).

It was announced on 17 July 2020 that the date of Scotland's next census would change from 22 March 2021 to 20 March 2022, due to the impact of COVID-19 on vital preparations for the census.

[2] See www.scotlandscensus.gov.uk/documents/Statistical%20Quality%20Assurance%20Strategy.pdf.

3. Introduction and Background

Not all people respond to the census. Therefore to yield an accurate estimate of the population size a separate survey called the Census Coverage Survey (CCS) is carried out. The CCS is designed to have independence from the census. For example the collection mode is different (the CCS collection is enumerator lead), and the address frame is produced manually. The records from the CCS are linked to the census records. As the census and CCS are independent, the number of linked and unlinked records can then be used, through dual-system estimation (DSE), to estimate the total population. This is the estimation[3] stage of the census processing (see Figure 1).

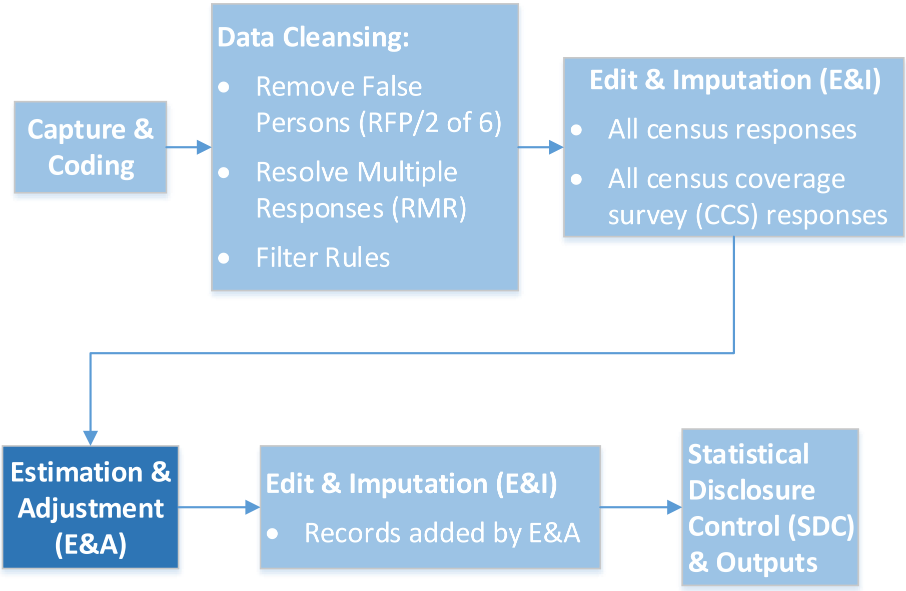
Capture & Coding
Data Cleansing:
Remove False Persons (RFP/2 of 6)
Resolve Multiple Responses (RMR)
Filter Rules
Edit & Imputation (E&I)
All census responses
All census coverage survey (CCS) responses
Estimation & Adjustment (E&A)
Edit & Imputation (E&I)
Records added by E&A
Statistical Disclosure Control (SDC) & Outputs

Figure 1 Where estimation fits into Statistical Data Processing.

An accurate population estimate relies heavily on accurate linking between the CCS and census records. The CCS is a sample of around 1.5 per cent of postcodes. Thus each match that is not linked will increase[4] the population estimate by around 80. Conversely each link that should not have been made will reduce the population estimate by around 80. Therefore the impact of a single error at this stage has a

[3] See NRS (2020) for more information on estimation in the 2022 Census.

[4] The DSE formula (NRS, 2020, p10) is $\hat{N} = \frac{N_{CCS}N_{Census}}{N_{CCS\cap Census}} \approx \frac{70{,}000\times 5{,}000{,}000}{66{,}000} = 5{,}303{,}030.3$. If a single match was not linked then this would become $\hat{N} \approx \frac{70{,}000\times 5{,}000{,}000}{65{,}999} = 5{,}303{,}110.7$, that is, 80.4 larger than the previous estimate.

much larger impact on the final census outputs than individual errors at other processing stages, such as Remove False Persons and Resolve Multiple Returns[5] (where a single error would typically affect the estimate by around 1).

[5] Information on these methodologies will be published on our website soon: External Methodology Assurance Panels (EMAPs) | Scotland's Census

4. 2011 Method

In 2011 the commercial linking software LinkPlus was used to identify individual-level links. These were then aggregated up to household-level links. Data for linking became available in April 2012 and was completed in September 2012. 100 person hours were required for clerical review, plus one member of staff working on it for a large part of the year.

Linking was done over five phases. These focused on links where both records were at the same location. After the LinkPlus phases, a manual search was done on the remaining unlinked CCS records. This was referred to as reconciliation.

The linking variables used were:

- First name
- Last name
- Date of birth
- Sex
- Address token (usually house name/number)

It was decided that the 2011 method would not be used in 2022 because it:

- Used commercial software which is no longer available to the admin data team
- Focussed on links between records at the same location, so may miss some links
- Relies heavily on manual searching for links
- Was unclear how (or if) links were chosen that did not need to be clerically reviewed

It was therefore decided to develop a new methodology, that could be audited better.

5. Proposed 2022 Method

5.1 Method Summary

To address the concerns mentioned in the previous section, the new methodology has the following features:

- It is written in SAS, which is readily available in NRS reducing the need for licences and training in other applications, and can be reviewed easily by other statisticians
- It contains 16 blocking[6] variables, six of which have no geographic element, allowing records to link from different locations (detecting cases where people have moved house, or respond in different locations for some other reason)
- Within the blocks every CCS record is compared directly with every census record
- By having comparatively wide blocks and making all possible comparisons within them reduces the risk of false negatives (missed matches)
- For each CCS record how similar it is to each census record in the same block is recorded, apart from where the difference is very large. Therefore for every CCS record an ordered list of possible matches can be presented. Reviewers can therefore peruse this list rather than performing a time-consuming manual search for links
- The new method is intended to be as thorough as possible, while still running in about two days when running on the CCS and census. There is a focus on the CCS records (as there are fewer of these) and then searching among the census records for links.

The following is an outline of the steps involved in the new method. These steps are discussed in detail in Section 5.2.

[6] When blocking, the records for linking are separated into blocks with the same value of some blocking variable(s). Links are only sought within (rather than between) blocks. There will then be no links where the linked records have different values for the blocking variable(s). See Steorts et al. (2014) for a discussion of blocking.

1. Standardise the identifiable data
2. Generate linking variables
3. For each of a range of blocking variables:
    a. Compare each CCS record to all census records that agree on the blocking variable, assigning scores for the similarity of the components (name, date of birth, sex and postcode)
    b. Categorise each link using the scores and assign distance scores
4. Bring together all the links, along with their categorizations
5. Find linked households
6. Search for further links among the linked households
7. Analyse the links to decide which to accept without review (there will be a sample check performed on these as part of the quality assurance process)
8. Find the CCS records not included in those accepted, and pass them for review, along with a list of linked census records ordered by similarity
9. Identify household links from the address and identified person links

The flow through the steps is summarised in the flow chart in Figure 2.

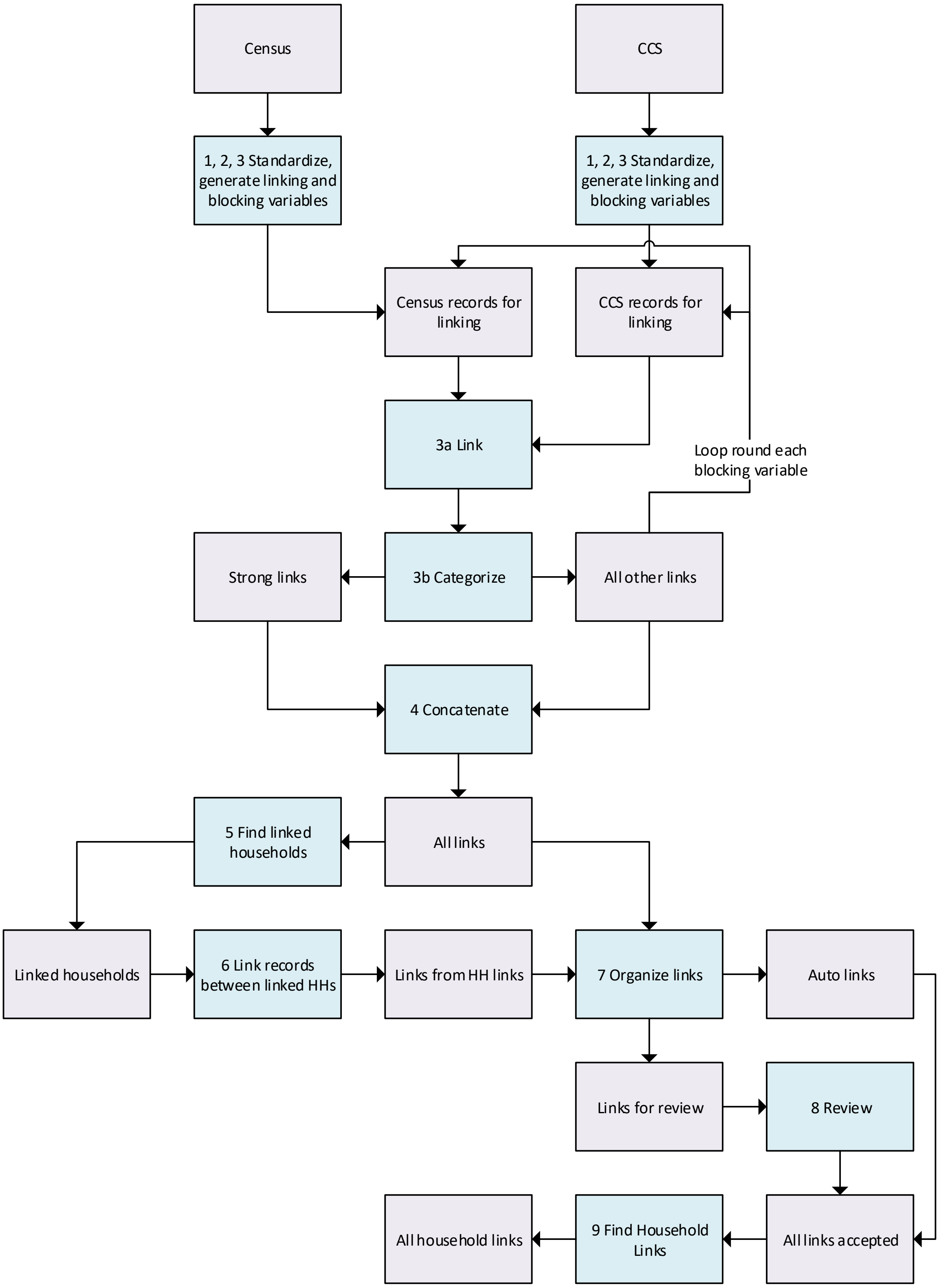

Figure 2 Flow of data (red boxes) and steps (blue boxes) through the process. The numbers indicate the number of the steps, as indicated in the text in sections 5.1 and 5.2.

5.2 Method Detail

Section 5.1 listed the broad steps for performing linking. These steps are now explored in detail using the same numbering as in Section 5.1.

Step 1: Standardise the identifiable data

The identifiable data is standardised to the following variables:

- First name[7]
- Middle name(s)[8]
- Last name[9]
- Day of birth[10]
- Month of birth
- Year of birth
- Sex[11]
- Postcode[12]
- Unique address[13]

Initial cleaning of the names involves removing accents from letters.

In addition a unique identifier is stored with each record. This allows tracking of individual records across datasets, and also allows records to be grouped together into households.

[7] First name is taken to be the first string in the first name field on the individual questionnaire. If this is missing then the corresponding field on the household questionnaire is used.
[8] The remainder of the first name field on the individual questionnaire (after first name has been extracted) is stored in middle. If this is missing then the corresponding field on the household questionnaire is used.
[9] Last name is taken from the last name field on the individual questionnaire, unless this is missing, in which case it is taken from the last string (space delimited) from first name field on the individual questionnaire. If this is missing then the corresponding field on the household questionnaire is used.
[10] Day, month and year of birth are taken from date of birth, and are stored as text strings of length two, two, and four respectively (with leading zeros).
[11] Sex is set to be 'M' for male, 'F' for female and missing for any other values (such as invalid).
[12] Postcode is stored with no spaces.
[13] In the 2011 test this was taken as a combination of postcode and property number (which itself may be a combination of flat and street number.

Step 2: Generate linking variables

At this stage a range of variables for use in linking are generated. These are derived from the eight variables generated at Step 1. From postcode the postcode area (for example, 'EH'), postcode district (for example, 'EH12') and postcode sector (for example, 'EH12 7') are derived and saved. In the 2011 test, when postcode is missing the area, district and sector are attempted to be derived from the CD and ED variables (which are components of the census unique identifier). These will only be set when all the records in the same CD/ED combination with non-missing postcodes have the same area, district and sector respectively. It is expected that there should be much fewer cases of missing postcode in 2022, but also that something similar will be possible in 2022.

A range of linking variables are derived from the name variables. Special characters are removed from names. However, before this is done the parts of the name that are delimited by special characters are extracted and stored in a variable for particular parts of the name. Other linking variables include phonetic encodings such as the Double Metaphone[14] and an extended version of this. For first name there are two nickname root variables. For example if first was either 'Alexander' or 'Sandy' then the nickname variable would be set to 'Alexander' to allow these two records to link. This variable also allows variant spellings of the same name to link (for example 'Jonathan' and 'Jonathon'). This will be particularly important when the CCS data is collected verbally and spellings may have been assumed rather than confirmed. Finally a fullname variable is generated from a concatenation of first, middle and last name variables (with spaces removed). The linking variables are used in the scoring, and so further information about them is included in Annex 2 on scoring.

Step 3: Blocking Variables

When linking is carried out each CCS record is compared with each census record that has the same value of a blocking variable as the CCS record does. This is

[14] The double metaphone was presented in Philips (2000).

repeated a number of times for different blocking variables. For some blocking variables there is an additional condition that some of the remaining variables must be missing on at least one of the records. The list of blocking variables, and their corresponding conditions, are given in Table 12 in Annex 1. These blocking variables were chosen so that for the vast majority of 2011 links, the census and CCS records would be in the same block for at least one of the blocking runs.

After the first two blocks (postcode and postcode sector) have been used, any records that link strongly and uniquely are removed from the set of records available to link. Such links would be expected to be accepted without review, in which case searching for weaker links would simply waste processing time.

Step 3a: Scoring

For each combination of CCS record and census record within the block, the similarity is measured and scored. Ultimately, this information feeds in to decisions on whether to automatically accept a link, and if it is not automatically then it is used to prioritise it in the clerical review list. The scoring is done separately for each of the components: first, middle and last names, sex, and date of birth. For each of these components there is a score to indicate the strength of evidence for this being a match, and another for the strength of the evidence against it being a match (that is, the strength of the evidence for a non-match). The scores from the different components are totalled separately for the evidence for and evidence against, to give a total score for the evidence for a match and the evidence for a non-match.

Table 1 Maximum possible for scores for first, middle and last names, date of birth, sex, and also when these are totalled. Links will get these scores if the variable is non-missing and identical in the two linked records.

Variable	Maximum For Score
First Name	50
Middle Name	25
Last Name	50
Date of birth	36
Sex	5
Total	166

Variables that can contribute strongly to the evidence for a match can get larger scores than those that do not (see Table 1). These scores were intended to reflect the relative importance for each variable for identifying matches, as judged through clerical review.

The detail on the scoring is described in annexes 2 and 3. The specific rules were developed in order to best replicate judgements from previous clerical reviews.

Table 2 Some example (mock) data for CCS and census. These data are referred to in the links in Table 3.

Source	Record	First	Middle	Last	DoB	Sex	Postcode
CCS	1	Davy	Kenneth	MacKenzie	1/1/1960	M	EH1 1AA
CCS	2	Maya		Patel	2/3/1950	F	EH1 1AA
CCS	3	Davina	Jane	Wilson	4/5/1970	F	G1 1AA
Census	1	David	Kenneth	MacKenzie	1/1/1960	M	KY1 1AA
Census	2	Maya		Patel	3/3/1950	F	EH1 1AA
Census	3	Davina		Wilson	4/5/1970	F	G1 1AA

To exemplify the linking, some (mock) example data is presented in Table 2 for the CCS and the census. Intuitively a reviewer could see that each of the CCS records matches the corresponding census record, even though in the first case the records are from different locations and has a nickname on the CCS, in the second case there is a slight difference in date of birth, and in the third case middle name is missing from the census record.

Table 3 For and against scores, distance scores, postcode scores and overall distance scores for links between the example (mock) data from Table 2.

CCS	Census	Evidence for a match						Evidence against a match					
record	record	First	Middle	Last	DoB	Sex	Total	First	Middle	Last	DoB	Sex	Total
1	1	20	25	50	36	5	136	0	0	0	0	0	0
1	2	0	0	0	0	0	0	15	0	35	12	5	67
2	2	50	13	50	12	5	130	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	14	47	18	5	84
3	3	50	0	50	36	5	141	0	0	0	0	0	0

In linking, each of the CCS records are compared with each of the census records if they are in the same block for any of the blocking variables. The results of this linking are shown in Table 3. The blocks that would be used for these five links would be date of birth, postcode, postcode, first four letters of first name, and postcode respectively. Each of these pairs of records are first compared on first name, middle name, last name, date of birth and sex. For each of these there is a score indicating the strength of evidence from the variable that the pair is a match, and a score indicating the strength of evidence against it being a match (see Table 3). These are summed to get total for and against scores.

Step 3b: Categorisation

Once the various scores have been calculated, each link is categorised. There are 22 different categories (which are listed in Annex 4). The categories organise the links into different types. The development of the categories was a manual process informed by clerical review. These group links according to the strength of the evidence for and against a match and also the type of relationship between the records. For example if the surname and date of birth are the same but the first name provides evidence of a non-match then this may be a set of twins, so such links are categorised as potential twins. The categorisation makes use of the for and against scores for each individual variable, and also of the for and against scores totalled across all the variables.

Table 4 Postcode score assigned to links depending on the level of agreement on address and postcode.

Agreement level	Postcode score
Address	1
Postcode (unit)	3
Postcode sector	4
Postcode district	5
Postcode area	6
One or both postcodes missing	7
No agreement on postcode	8

Each category is given a distance score. The distance scores range from zero (exact agreement) to nine (most likely a non-match). The distance scores are used to group links according to how strong the evidence is that they are a match. Any link with a distance score of nine, that is, a link that can confidently be considered a non-match, is deleted at this stage. (However, none of the links made using the postcode blocking variable are deleted, even if their distance score is nine.) In addition a postcode distance score is calculated, based on the extent to which the postcodes of the two records agree (see Table 4). The postcode distance score and the distance score from the categories are then summed to get a total distance score.

Table 5 Distance scores, postcode scores and overall distance scores for links between the example (mock) data from Table 2.

CCS record	Census record	Distance score	Postcode score	Overall distance score
1	1	1	8	10
1	2	9	3	12
2	2	1	3	4
3	1	8	8	16
3	3	1	3	4

Table 5 shows the distance scores obtained for the mock examples from Table 2. The categorisation is done using the scores from Table 4. Each category is associated with a distance score from 0 (exact agreement) to 9 (completely different). These distance scores are then added to the postcode score to give an overall distance score. These overall distance scores are what is primarily used to organise the links.

In moving from the continuous for and against scores to the 10 distance discrete scores, some detail is lost. However there are a number of advantages to this. The first is that it allows us to make human judgements of how the scores from the different variables combine. For example a pair of records with high overall for scores might be twins if the first name is different, or a parent–child pair if the year of birth is sufficiently different. However, links with similar overall for and against scores might more plausibly be the same person if they have a different pattern. The second reason is that it makes it easier to combine this score with similarity information from the postcode. Having an overall combined score allows for the links to be ranked from best to worst (strongest to weakest). However, one of the most important reasons is that it allows us to count how many links there are of particular overall distance scores. This will become important at Step 7 below.

Further detail on the categories is included in Annex 4.

Step 4: Concatenate results from different blocking variables

Steps 3a and 3b are repeated for each of the blocking variables. For each blocking variable a dataset of categorised links is produced. At Step 4 these datasets are combined. Specifically, these are interleaved on the CCS record ID, overall distance score and (category) distance score. This ensures that all the links found for a particular CCS record are together, and that the strongest such links appear first.

Step 5: Identify linked households across postcodes

This step considers the households of persons that have been linked. This is so that, at the following step, other person records in these household pairs can be compared to find more person links. Such links may have been missed if the person records were not in the same block for any of the blocking variables. This could happen if a household has moved house between the times of the census and CCS. The detail of this step is below.

The full list of links is then searched for the best links for each CCS record. For each of these the best distance score and the best overall distance score (the latter including the postcode score) is noted. If the best distance score is six or better and the best overall distance score is 12 or better, then the CCS record ID is noted along with its best distance and overall distance scores.

For the CCS records thus identified all the links that have distance scores and overall distance scores are as good as the best such scores for the CCS record are found. For all these links (apart from links where the two records are in the same postcode) a note is made of the CCS and census household IDs and these are saved in a dataset of linked households.

Step 6: Link all the records between the linked households

For all the linked households the individual records in the CCS household are compared with the individual records in the linked census household, in the same way as in Step 3. This gives another chance to make links. These links are then appended to the overall list of links.

Step 7: Analyse links and accept some without review

For each CCS record the number of links at each overall distance score is counted. If there is only one link at the best overall distance score (as long as that score is three or better) then that link is accepted without clerical review. At overall distance scores from four to seven there is an additional condition that there are no links at overall distance scores immediately adjacent. All remaining cases will be passed for clerical review. In addition, a sample of the links indicated for acceptance without review will be clerically reviewed, in order to quality assure the process.

Step 8: Pass remaining CCS records, along with a list of potential census matches, for review

The CCS records that are not automatically accepted (those whose best overall distance score is eight or greater, and those whose best overall distance score is not uniquely good) are then passed for review. These are presented along with the census records they were linked to. These census records are sorted so that the census records that are most likely to be a match for the CCS record appear nearest the top of the list. In this way, if a match exists for the CCS record, then the reviewer should generally be able to find it fairly quickly. The census records are therefore sorted by the following variables:

- Whether another CCS record in the same household linked to another census record in the same household sufficiently strongly to be accepted without review
- Overall distance score
- The total evidence for a match score minus the total evidence against score

- The total evidence for a match score

This information will then be loaded into a file within the secure NRS IT area. In the file there is a view that loads in information for a particular CCS record and the linked census records. The identifiable information for the census records will be shown alongside the relevant information for the CCS record (that is, there would be a column for first name, last name, postcode, and so on). In addition, differences and similarities will be automatically highlighted through shading. There will be options to allow reviewers to access further information about the CCS record or particular census records. This would include information about other individuals in the same household.

When a reviewer has reached a decision about a link they will be able to record this. This will record whether the CCS record was considered to not have a matching census record, or if it did then which census record was considered to be the match. Once this information has been recorded the information for the next CCS record and its associated census records would automatically be loaded, and the process would be repeated. As part of quality assurance, samples of these clerical reviews will be assessed to ensure consistency between reviewers.

Step 9: Link responding households using address and person links

While the census–CCS person links will be used to estimate the total number of people in Scotland, a set of household links is needed to estimate the number of households. The census definition of a household is:

> *One person living alone, or*
>
> *A group of people (not necessarily related) living at the same address who share cooking facilities and share a living room or sitting room or dining area.*

This definition is taken to mean that a household is both the people **and** the address. Therefore, a household link will be formed between census and CCS if a return is received for the same address (that is, an address that is matched) in both the census and CCS, with at least one corresponding person link. In most cases the linked household returns will relate to the same people. However in some cases people may have moved out of an address following census day, and other people may have moved in before the CCS. In these cases the household should only be linked if at least one person record is linked between the two, in spite of the address being the same.

The census–CCS person linking is intended to identify what proportion of people did not respond to the census by identifying people who appear on the CCS but not the census. Some people who appear on the CCS should not have appeared on the census, and so are considered out of scope. This will include babies born since census day, and people who were not living in Scotland on census day. People who were not living in the CCS area on census day could also be considered out of scope as what is required is only the sample of the CCS area, which can then be scaled up across Scotland.

Occasionally people will move into or out of CCS areas between the two surveys. This could cause them to be missed from the CCS, or to be counted at a location other than where they were on census day. Persons on the CCS who appear on the census at a non-CCS area will be counted as out of scope for the purpose of the DSE calculation. However, these links are used in a separate exercise dealing with overcount correction.

There could be similar issues with household links. As we define households as a single person, or group of people living at the same address, we would not want to link two different groups of people who did not live together at an address. Any household which became occupied between Census and CCS would need to be identified, otherwise they would be considered an unlinked CCS household and cause overestimation. Excluding households in the CCS where none of the residents lived there on Census day from the DSE calculation avoids introducing this systematic error, and also avoids complications with person records needing to be out of scope while the household itself is not.

Methodology for address linking is currently being developed and will be described in a forthcoming paper. There may be CCS households at addresses that cannot be linked to the census returns. For these households a household link could be sought by considering the person links of the people in that household. These links could then be reviewed to determine whether they are likely to be the same address. If they do not appear to be the same address, the link should not be included in DSE.

6. Results Using 2011 Data

6.1 Testing plan

Under the method above there are a set of links that should be accepted without review. Therefore, part of the test is to ascertain whether any of these links are bad links. A 'bad' link is one that has been clerically reviewed and where the reviewer felt that it was likely a non-match (that is, the records represented different persons). Conversely, a 'good' link is one that has either been clerically reviewed and the reviewer felt that it was a match, or it was a link that was used in 2011.

For each other CCS record it is proposed that links are sought from a the list of suggested census records. Ideally, matching records should appear at the top of the list. Therefore, a measure of the quality of the method would be how close to the top the list any good links are. Good links further down the list should still be found by clerical reviewers, as the whole list will be available to them. However, in some cases there can be hundreds of such records presented, so reviewers could not reasonably be expected to check all of these for each CCS record. If good links are too far down the list then they may be missed, resulting in a false negative. It is difficult to say how likely good links are to be missed as a function of how far down the list they appear. For the purposes of producing a measure on this, the number of links that appear 20th or lower in the list will be considered.

The three measures of the success of the linking method are therefore (with lower values indicating better performance):

- The number of bad links that are accepted without review
- The number of good links that are not detected by the method
- The number of good links that are detected but are placed 20th or beyond in the list of suggestions

In deciding whether a link is a good link or a bad link, consideration is made of whether the link was one of those accepted in 2011. 67,653 links were made in 2011 (94.1 per cent of the 71,898 CCS records). However only 66,881 of these links are known, as some of the 67,653 were considered out of scope and so not used.

For this test these 66,881 2011 links are considered good links. In addition, 3,004 links were found by the admin data linking methodology and clerically reviewed in previous exercises. Those links were evaluated as yes, no or maybe (being matches). The 1,195 links classed as yes are considered good links, while the 1,071 links classed as no are considered bad links. The 738 maybe links, along with any links that were not used in 2011, nor evaluated in the previous exercise, are clerically reviewed.

In addition, the method should be as efficient as possible (without compromising quality). The two relevant[15] factors are therefore the processing time and the amount of clerical review. The amount of clerical review that is required will depend on how many links can be accepted without review. It would be undesirable for the method to be over conservative and pass the burden of assessing links on to manual reviewers. That could delay the processing and could also affect quality as a higher level of consistency should be possible if the links can be assessed automatically. Finally the review process should be faster if the good link appears high on the list of suggestions, preferably first, as this would save the reviewer time searching through the list.

Therefore the efficiency measures are:

- The amount of processing time the method takes
- The number of CCS records that are passed for review
- The distribution of the good links in the list of links to review

6.2 Bad links incorrectly accepted

Table 6 shows the number of links passed for automatic acceptance by the overall distance score and the classification. In total 65,518 links (91.1 per cent of the 71,898 CCS records) were passed for automatic acceptance. Encouragingly, the vast majority of these links (over 99.9 per cent) are already considered good links,

[15] The time to develop the method is less if a factor as this can be carried out in advance of census collection, and so will not impact on when the first outputs can be released.

especially at the stronger distance scores. No bad links were included. There were 10 links that were classed as maybes and 51 that were previously unknown.

Table 6 Links that are accepted without review by overall distance and previous assessment. Those indicated as original 2011 links are those links that are known to have been used in 2011, other links may have been identified in 2011 but were not used because they were considered out of scope. Links that were previously reviewed following previous linking exercises are categorised as yes, no or maybe, depending on the assessment of the reviewer. Any previously unknown links are listed as unknown.

Overall distance score	Original 2011 links	Yes	Maybe	No	Unknown	Total
1	6,082	1	0	0	0	6,083
2	38,018	11	0	0	6	38,035
3	3,891	4	0	0	5	3,900
4	13,134	57	0	0	14	13,205
5	2,071	183	1	0	13	2,268
6	929	195	2	0	6	1,132
7	587	294	7	0	7	895
Total	64,712	745	10	0	51	65,518
Percentage	98.8	1.1	0.0	0.0	0.1	100.0

The 10 maybe links and the 51 unknown links were then individually reviewed. All but 11 of these links were considered good links. Further investigation will explore why these were automatically accepted and amendments made to ensure that such links are passed for review.

6.3 Good links missed

Three of the original 2011 links and nine of the yes links were not detected by the method. These cases were generally when there was a slight difference in each linking component (name, date of birth, postcode). Taken as a whole these differences are small enough that they do not suggest that the link is bad, but with none of the components being exact means that the records do not agree on any of the blocking variables so the link is not tested. Missing 12 matches would increase the population estimate by around 900, over-coverage of around 0.017 per cent. This error is smaller than the Key Performance Indicator for the bias (0.5 per cent).

6.4 Good links placed low in the list

6,816 CCS records were passed to review[16]. 2,990 good links (2,589 original 2011 links and 401 yes links) are included in the links to be reviewed. Of these 44 (34 original 2011 links and 10 yes links) appeared 20th or later in the list of links to review for the particular CCS record (affecting 0.6 per cent of CCS records passed to review, 0.06 per cent of all CCS records). If these were missed then that would result in over-coverage of around 3,500 (around 0.07 per cent).

It may be possible to reduce this. Sometimes the good link has been pushed down the list by a group of links that are very similar (perhaps when one of the linking components was missing). In such cases it may be possible to group these records together so that the reviewer could make a decision about them as a group. That is, if the reviewer decides that one of these records is not a match, and there are a group of records that are very similar, then this would suggest that none in the group should be considered a match. This would allow the reviewer to get down to records further down the list more quickly, potentially finding these good links.

6,816 CCS records were passed for review, each with a list of census records. It may be possible to slightly reduce this number slightly with better discrimination between good and bad links. However it should be noted that 56 per cent of these CCS records have no known good links. Thus while theoretically some of these CCS records could be removed from the set of those to review, the majority of them would need to be reviewed to check they did not link, however good the linking method was.

As most of the reviewed CCS records do not have good links the reviewer would need to look through a number of possible links in order to be sure that there was no good link. This may take up to around a minute, so for estimation purposes two minutes is used to account for variation in speed. Assuming a seven-hour day this is equivalent to 32 person working days. It is not currently known how many reviewers

[16] A further two CCS records were not linked automatically or passed for review. This was because they had no information for name, date of birth or postcode, and so could not be linked to any census records.

will be available, but this suggests that a clerical review team of 10 people could complete the clerical review in around three days. Added to the around two days processing time, this suggests that a concerted effort could result in the CCS linking being completed within a working week, assuming no major problems.

The reviewing process can be made easier if the good links tend to appear early in the list of possible links. Table 7 shows where in the list the good links would appear. Encouragingly most of these (87 per cent) appear top of the list, with about half of the remainder appearing second in the list.

Table 7 Number of good links by the place they appear in the list of links to be reviewed.

Position in list	Number of good links	Percentage of good links
1	2,594	86.8
2	191	6.4
3	76	2.5
4	29	1.0
5–9	41	1.4
10–14	6	0.2
15–19	9	0.3
20+	44	1.5
Total	2,990	100.0

6.5 Household linking

To test the household linking, the address links and the person links between the 2011 census and CCS were considered. The person links included all the original 2011 person links, and the links that were found later and reviewed as 'Good' links. This resulted in 68,076 person links.

These person links represent 30,823 household links. In counting household links multiple links are counted multiple times. For example if person 1 and 2 are in household A in the CCS, and in households B and C respectively in the census this would count as two household links.

The CCS households were also linked to the census using the address. This was done on postcode and the first part of the address (a variable called enum_name).

In some cases the address does not agree exactly, because the address is recorded in different formats in the census and the CCS. In 2022, it is expected that address information will be of better quality, especially on the census, where respondents will select their address from a list.

For each CCS household, there may be a link to the census using the address or person based methods. For the CCS households that have a link using both methods, a comparison can be made to determine whether these links are to the same census household. In cases where persons in a CCS household linked to multiple census households just one link was counted for each CCS household. This prioritised the links that were the same as the address-based link, and otherwise the household link that involved the largest number of people was taken. The results from this are shown in Table 8.

Table 8 CCS households by whether they link to a census household using the person-based and address based methods. Where a household links by both methods it is indicated whether the linked census household is the same in both cases.

Category	Number	Percentage
HH same for person and address base	22,148	68.9
HH different for person and address base	397	1.2
Person-based only	8,010	24.9
Address-based only	126	0.4
Neither	1,441	4.5
Total	32,122	100.0

It can be seen from Table 8 that 8,010 CCS households (25 per cent) link to the census on the person-based but not address-based methods. However, in some of these cases the address may be the same, but just recorded differently in the census and CCS.

To explore whether this is the case, Table 9 breaks down the number of household links where there was only a person-based link, or where the person-based and address-based links were different, by whether the person-based link had the same postcode for the census and CCS record, or whether these were different. It can be seen that most of these do indeed have the same postcode. A visual inspection of these links suggests that the address is indeed generally the same, but differences in recording or scanning errors simply prevented them from being linked exactly.

In addition, Table 9 breaks down the CCS cases by whether the same link was used in 2011, whether a different link was used, or whether the CCS household was not linked in 2011. Encouragingly, most of the household links found are the same as those used in 2011. In some cases the CCS record was linked to a different census household, especially in cases where the CCS household links to different census households using the person and address based linking. 649 CCS households (2,090 – 1,441) were linked to census households that were not linked in 2011. None of the CCS households that were linked in 2011 were not linked in the current run.

Table 9 CCS household links comparison with household links used in 2011 (that is, whether the link found was the same as that used in 2011, different from it or if the CCS household was unlinked in 2011).

Person-based and address-based household link	Comparison with 2011 link			Total
	Same	Different	Unlinked	
Same	22,144	4	3	22,148
Different: postcode same	1	234	3	238
Different: postcode different	8	12	139	159
Person-based only: postcode same	7,536	0	10	7,545
Person-based only: postcode different	76	21	368	465
Address-based only	0	0	126	126
Neither	0	0	1,441	1,441
Total	29,764	268	2,090	32,122

This suggests that using the rule that a household link should be at the same address and have at least one person link, would identify the vast majority of the household links as were used in 2011 (around 22,144 + 7,536 = 29,680, 99 per cent of the 30,081 2011 household links). It is anticipated that the address linking will therefore need to use some inexact linking. In addition, clerical review may be needed for unlinked CCS households, especially where the persons in the household linked to census records in the same postcode. The detail for this process is still to be fully determined.

7. Results Using Rehearsal Data

7.1 Person Linking

In October 2019 a rehearsal of the census was conducted in parts of Glasgow, Dumfries and Galloway, and the Western Isles. The method was tested again using the 2019 rehearsal data. The rehearsal data was linked to a synthetic CCS, which consisted of NHSCR[17] records in a sample of postcodes in the rehearsal areas. This synthetic CCS contained 3,132 records.

Table 10 Number of synthetic CCS records and automatically accepted links by rehearsal area. Number of links also shown as a proportion of the synthetic CCS records.

	Postcode area			Total
	Glasgow	D&G	Western Isles	
Synthetic CCS records	1,499	854	779	3,132
Automatically accepted links	154	278	282	714
Links/synthetic CCS records	0.103	0.326	0.362	0.228

Table 10 shows the proportion of synthetic CCS links that linked to the rehearsal strongly enough to be automatically accepted. It can be seen that this is much lower than in the 2011 test (overall 22.8 per cent in the rehearsal, compared with 90.5 per cent in 2011). However, this is to be expected, as the census rehearsal had a return rate of 25 per cent[18], much lower than that for the 2011 census. Therefore, many of the people who appear on the NHSCR would likely not have responded to the census rehearsal, and so a corresponding record would not be available for these records.

To avoid this problem, Table 11 shows the automatically accepted links as a proportion of the rehearsal records. This proportion (86.2 per cent) is now much closer to the proportion from 2011. It remains somewhat lower, especially in Glasgow. This may be because of differences between the reference dates of the two datasets (the rehearsal is 13th of October, while the NHSCR is the 30th of June).

[17] National Health Service Central Register, an administrative dataset of people registered with an NHS GP.
[18] See https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20Scotland%27s%20Census%20Rehearsal%202019%20-%20Evaluation%20Report.pdf.

This may explain why Glasgow is particularly affected; urban areas may have more people moving house over this timescale, especially in a student area where the reference dates are in different academic years.

Table 11 Number of rehearsal records in synthetic CCS areas and automatically accepted links by rehearsal area. Number of links also shown as a proportion of the rehearsal records.

	Postcode area			Total
	Glasgow	D&G	Western Isles	
Rehearsal records in CCS areas	204	317	307	828
Automatically accepted links	154	278	282	714
Links/rehearsal records	0.755	0.877	0.919	0.862

All the automatically accepted links were clerically reviewed. All were considered to be matches with the exception of two. These two links were where the date of birth and postcode differed slightly, and the name, while not especially common nationally, was common to the local area. As a result we plan to implement a process that will measure how common names are in each area (likely postcode district). This can then be used to assess the plausibility of two records representing different persons. If that plausibility exceeds some threshold then the link will then be passed to clerical review instead of being automatically accepted.

The remaining synthetic CCS records were also clerically reviewed against the list of census rehearsal records that were considered potential matches by the algorithm. Among these a further 32 links were found that were considered matches. It is encouraging that this is low, as it suggests that the vast majority of genuine matches are being automatically linked by the algorithm.

From rehearsal the possibility of comparing links to a third source was raised. If a linked census record linked to a particular record in the third source, and the CCS record in the same link linked to a different record in the third source, then this might indicate that the link was a non-match.

To test this, records from the links from rehearsal that were flagged for automatic acceptance were linked to the 2011 census. This was done blocked on postcode and date of birth. In only one case did the linked rehearsal and synthetic CCS records link to different 2011 census records. Encouragingly, this was one of the two links that was considered a non-match in clerical review. We are therefore considering linking the 2011 census dataset to the records in the census–links to help inform decisions around whether they are matches or non-matches. Further work will be done to decide what blocking variables should be used.

Just to be completely clear: no data from the 2011 census would be used in the 2022 census results. This is purely to assist decisions around wither particular links between the 2022 census and 2022 CCS are matches or non-matches.

7.2 Household Linking

The synthetic CCS was built using NHSCR data, which does not include addresses. As such the synthetic CCS data cannot be organised into households. Therefore, the household linking cannot be tested using the rehearsal data.

8. Strengths and Limitations

When running on the 2011 census and CCS the process takes around two days to run. This includes time to load the data and generate the linking variables. Generating the linking variables for the census records takes around 38 minutes. Performing this step in advance of the CCS data being available could save processing time when linking. Furthermore, running this in batches will allow clerical review to begin before the whole dataset has been linked.

Most of the links that are ultimately accepted are passed for automatic acceptance. This helps reduce the amount of clerical review time needed. Very few of these links were considered non-matches (only two in rehearsal and 11 on the 2011 data). However, we are pursuing methods that will help eliminate such false positives. This includes considering the uniqueness of linking variable values in local areas (that is, to account for the locally common names issue) and also linking to the census 2011 dataset to see if the link represents two persons who were recorded on the census in 2011.

The proposed method effectively combines the clerical review and manual searching steps from 2011, which should reduce the amount of time needed for manual linking. By limiting the search to a specific list of census records for each CCS record, ones that agree on one of the blocking variables, there is a risk that some matches are missed. However, it would be unfeasible for a reviewer to search through all 5,000,000 census records looking for a match for an unlinked CCS record. Such a manual search would therefore need to filter the census records down to a manageable set, presumably those that agreed on some variables, or part of variables. Such a process would then likely reduce to something similar to the planned clerical review process, and so the risk of false negatives from a manual search would likely be at least as high as in the proposed process.

The method is able to link records from different locations. It is important to know when people appear in different locations on the census and CCS. Work is also underway to develop address linking methods, which should help link addresses that are recorded slightly differently. Use will also be made of a CCS question that asks respondents where they were on census day if that differed from the address they responded to the CCS at. By using both the CCS address and this alternative address will give different options for finding links, and will strengthen the evidence when links are made between locations.

Also, by developing the method in house, using a commonly used language, makes the method more controllable.

9. Conclusion

The linking method presented here links the CCS to census in around two days. The vast majority of known good links were either automatically accepted or should be detected through review. Those that were missed would result in a systematic error in the population estimate of around 0.08 per cent. The method produces a list of possible links for each CCS record, eliminating the need for manual searches on the dataset (reconciliation). The time needed to complete the CCS linking exercise will depend on the number of reviewers available. It is not currently known how many reviewers will be available, but a team of around 10 reviewers could, all going well, complete the CCS linking exercise using this method in around a week. This is important in order to achieve the release of first census outputs within a year of collection.

10. References

National Records of Scotland (2019), *Statistical Quality Assurance Strategy*, (online) available at: https://www.scotlandscensus.gov.uk/documents/Statistical%20Quality%20Assurance%20Strategy.pdf

National Records of Scotland (2020), *PMP001: Estimation and Adjustment Methodology*, (online) available at: https://www.scotlandscensus.gov.uk/documents/Scotland%E2%80%99s%20Census%202022%20-%20PMP001%20-%20Estimation%20and%20Adjustment%20Methodology%20(pdf).pdf

Philips, L., 2000, 'The double metaphone search algorithm', *C/C++ Users Journal*, vol. 18, no. 6, pp. 38–43

Steorts, R., Ventura, S., Sadinle, M. and Fienberg, S. (2014) 'A Comparison of Blocking Methods for Record Linkage' in: Domingo-Ferrer J. (ed.) *Privacy in Statistical Databases: Lecture Notes in Computer Science*, vol. 8744, pp. 253–268

Zhao, C. and Sahni, S. (2019) 'String correction using the Damerau-Levenshtein distance', *BMC Bioinformatics*, vol. 20, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6551241/

Annex 1: Blocking Variables

Table 12 lists the combinations of variables that are used for blocking. In addition, any conditions that are placed when selecting links to assess are also given.

Table 12 Blocking variables used for linking, and any additional conditions.

Variable, or combination of variables	Condition
Postcode	
Postcode sector	
Birthday (day and month of birth)	
First name and postcode area	
First two characters of Double Metaphone of first name and of last name	
First five letters of last name	
First four letters of first name	Postcode missing on one or both records
First nickname and postcode district	
First nickname, first character of Double Metaphone of last name and postcode area	
First nickname and first two characters of Double Metaphone of last name	
First nickname 2 and postcode district	
First nickname 2, first character of Double Metaphone of last name and postcode area	
First nickname 2 and first two characters of Double Metaphone of last name	
First character of Double Metaphone of first name, first two characters of Double Metaphone of last name and postcode area	
First three characters of Double Metaphone of first name and postcode area	
First four characters of Double Metaphone of last name and postcode area	

Annex 2: Scoring of Name Comparisons

This section discusses in detail how the for scores (which indicate the strength of evidence for two records representing the same person) and the against scores (which indicate the strength of evidence for two records representing the different persons) are calculated for the various components. There are a number of attempts to find evidence for a match. Each one will update the for and against scores only if that will strengthen the evidence for a match.

Missing Names

If name is missing on one or both records then the for and against scores are both 0. Otherwise if a name component is exactly the same between the two records then the for score is 50 (25 for middle name) and the against score is 0.

For first names there is also a check for the name being 'BABY' on both records. In this case the for and against scores are both set to 0 as the guidance (in 2011) indicated that unnamed infants should be recorded as 'BABY'.

Nicknames

Another check for first names is nicknames. Thus if we had 'Alexander' on one record and 'Sandy' on the other then it is quite plausible that these are the same person, even though the first name strings are quite different. To perform this check we make use of the nickname linking variable. That variable is set to a particular value for a range of names that have the same nickname. Thus if first was either 'Alexander' or 'Sandy' (or 'Alex', 'Xander', and others) then the nickname variable is set to 'Alexander'. (The name groupings were built up manually, assisted by exploring links between datasets where last name, date of birth and postcode agreed, but first name did not.) Thus if the first names differ between records but the nicknames agree then the against score is set to 0 and the for score is set to 20. Some of these are specific to a particular sex. Thus if the first name is 'Alex' then the nickname will be set to 'Alexander' if sex is male and 'Alexandra' if sex if female.

There is also a second nickname variable that groups together more tenuous name groupings such as 'John' and 'Ian', which results in a for score of 10.

The nickname check also detects alternate spellings of the same name, such as 'Nicholas' and 'Nicolas'. This may be particularly important for Census Coverage Survey linking when data is reported verbally and spellings may not be confirmed. In total there are 189 groupings defined, and 45 more tenuous ones.

Character comparison for names

If none of these situations hold then the name components in the two records are compared at the character level using a method inspired by the Damerau–Levenshtein edit distance[19]. The characters in the name from one record are linked to those in the name from the other record. This is done by first comparing the characters at the same location in the strings. If these do not agree then this moves to adjacent letters, and then letters at a distance of two, and so on. Once this has completed there is a tidying up stage to ensure that adjacent letters are linked to letters at the same distance if possible.

Once the letters have been linked they are then analysed in order to identify the substitutions, transpositions, deletions, insertions and jumps would be required to transform one string into another. For each of these there is an associated score. These scores depend on the letters involved. For example if we need to insert a 'W' then that would attract a larger penalty than if we only need to insert a 'I' because a mark on a page may be mistaken for an 'I' in scanning, but is unlikely to be mistaken for a 'W'. Similarly for substitutions some changes are more plausible than others. Combinations like 'U' and 'V' can be easily confused, as can 'O' and 'D'. In total 50 such combinations are noted.

The scores from all the individual differences are then combined to give an overall score. That score is then converted to scores for and against the records being a match.

[19] See Zhao and Sahni (2019) and references therein.

Swapped first and last names

Sometimes people enter their names in an unexpected order. To account for this a comparison is made between the first name of one record and the last name on the other record and vice versa. If these both agree then the for scores for both first and last names are set to 40. If only one of these agrees then one of these scores is set to 40, while the other is set by doing the character comparison on the differing values. That is, if first_1 agrees with last_2 then the first for score will be 40, while the last for score will be set by doing a character comparison between first_2 and last_1.

Titles

If first name begins 'MR ' or 'MRS ' then that part is removed from the first name and stored in a variable called title. If the two records being compared both have 'MR' and 'MRS' respectively in their title variables, and their sex agrees with this information, then a penalty of 20 is combined with the for and against scores for first name.

Comparison to middle name

Some people go by what is officially their middle name. In order to successfully link these cases the first name for one record is compared with the middle name of the other. If this agrees then the for score for first name is set to 15 (unless it was already over 15). A similar check is also done between last name and middle name.

Compare name parts

Some people have double-barrelled first or last names. However they may go by only part of this. For example 'Sarah-Jane' may go by Sarah, or even Jane. To detect such cases we make use of other linking variables that pull out parts of names that are delimited by special characters. If these agree with the name from the other record then the for score is set to 25 (unless it was already over 25). This is done for first names and also for last names. In other comparisons special characters (including spaces) are removed before the comparison is made.

Comparing first letters of name or Double Metaphone code

The next check is to count the number of letters that agree at the start of the name from the two records. If so then the for score is set to be that given in Table 13. This covers a range from one letter agreeing to five (or more) letters agreeing. If only one letter agrees then this is treated differently, so that this method is used only if one record only has the initial (for example, if one record had 'Peter' and the other had 'P', but not if the other was 'Paul'). These scores are only used if they result in a higher for score than would otherwise be. Another exception is when 3 or fewer letters agree and the names are distinct but common. For example if we had Mary and Margaret then the first three letters agree, but as the names are common then this is not used to score the similarity.

Table 13 The for scores assigned when the first part of the name agrees either on the name itself, or the Double Metaphone coding of it. If only one letter agrees then this method is only used if one of the records only has one letter. * When only 1 letter agrees on name then this is only used if one of the names only has one letter.

Number of	Score when characters agree in:	
characters agreeing	Name	Double Metaphone of name
5+	20	20
4	13	13
3	7	9
2	3	4
1*	10	-

Similarly the first characters of the Double Metaphone[20] are compared. The Double Metaphone is a phonetic code, so this allows for detection of cases where a name has been written differently, but sounds the same. This is another situation that may be particularly common for verbally reported data such as the Census Coverage Survey. As a character in the Double Metaphone code can relate to more than one letter in the original string, agreement on Double Metaphone can indicate stronger agreement than agreement with the same number of letters on the original string. Therefore these scores are slightly larger than the equivalents for the agreeing letters on the original name.

[20] The double metaphone was presented in Philips (2000).

There is an exception when comparing the last names on the original string or Double Metaphone. If the last name begins ‘Mc’ or ‘Mac’ then the count of the agreeing characters is reduced by 2 and 3 respectively. This is because names beginning this way are so common, while being very distinct. Therefore we would not want to say that MacDonald and MacPherson were as similar as Scalon and Scanlan.

Full name

Sometimes a space is missing between the first and middle name, meaning that the middle name gets concatenated onto the first name. Other times a space gets inserted between letters of the first name, meaning that part of the first name gets put as the middle name. Another issue is that the whole name can be entered in the first name field.

All these issues can be resolved by considering the full name, that is, the concatenation of first, middle and last names (with spaces and other special characters removed). This full name is one of the linking variables used. It is compared between the two records. If it is not exactly the same then a character comparison is done. This allows a for and against score to be calculated for the full name. If this score is better than the for scores for first and last name then the first and last for scores are amended using the full name for score.

Annex 3: Scoring of Sex and Date of Birth

Sex

If sex is missing on either record then the for and against scores are both zero. Otherwise if sex is the same then the for score is 5, while against score is 5 if the sex is different.

Date of Birth

If the day, month and year components either agree between the records, or are missing on one of the records, then we count the number of these components were at least one of the records is has missing information. The for score is then given by: $12(3-m)$, where m is the number of components that are missing on at least one of the records. The against score is 0 in such cases.

If the dates of birth are non-missing on both records, the years agree and the day and month agree with the month and day on the other record then the for score is 20 and the against score is 0. This is to account for cases where the date has been entered in American format on one of the records.

Table 14 Sets of digits that may be confused in scanning, and so are given a smaller difference penalty.

Set of digits
2, 4, 5
8, 9
1, 7
3, 5, 8
2, 7
2, 3
5, 6
7, 9

If the two dates of birth are complete then the individual digits are compared. That is, the first digit of the day of birth from one record is compared with the first digit of the day of birth from the other record, then the second digit and so on. If the two digits are both in one of the sets given in Table 14 then we count this as a difference of 1. All other differences are counted as a difference of 2. (The particular sets of digits are chosen to be those that are often confused in scanning, so are more likely to be the same than for other pairs of digits.) These differences are then totalled across the whole date of birth.

There is an exception for the century. If this differs between the records then it gets counted as a difference of 2, rather than comparing each digit. This is because people sometimes confuse the century in the year if they are used to writing, for example, 19-- instead of 20--.

Another exception is if a digit appears in a different position in the component. For example if day was 21 on one record and 02 on the other then it may just be that the '1' was missed on one side and a leading zero added. Such cases when one record has a leading zero would then get counted as a difference of 2, rather than 4.

The totalled differences (*d*) are then put into the following formula: $6(3 - d - 2m)$. If this is positive then it is used for the for score (with against score being 0), and if it is negative then the for score is 0 and the against score is the absolute value of the formula.

A final check is to count the number of components (day, month and year) that are different. If only one is different, then the against score is set to 0.

Annex 4: Categorization of Links

Once the for and against scores have been calculated for each component for each link, the links are placed into one of the categories shown in Table 15.

Table 15 List of categories used to class the links along with a brief description of the condition used to place them and the distance score associated with the category. The categories are presented in order of the priority in which they are assigned. That is, links are only assigned to a given category if they do not meet the conditions for any preceding categories.

Distance Score	Name	Description of Condition
0	Exact	All components agree exactly and non-missing
7	Different – parent-child	Age difference ≥15, first and last for >0
6	Different – twin	Last for >15, DoB for >0 no evidence of match from first name
1	Same	Fairly strong evidence for match from first, last and DoB, no evidence against from sex or middle name
2	Same 2	As Same, but slightly weaker evidence
2	Goes by middle name	DoB, last and sex agree exactly and non-missing, first from one record agrees exactly with middle from other
4	Likely same (A)	Total for >70, total against =0, total for – last for >20
4	Female last diff	Female, fairly strong evidence for match from first and DoB, and last against >0
5	Non-female last diff	As Female last diff but without condition on being female
5	DoB same, miss name	DoB for >10, age difference <14, name missing on one record
4	Name same, miss DoB	First for ≥20 and last for ≥20 and total for >50, DoB missing on one record
5	Likely same (B)	Total for >45, total against =0, total for > last for + 15
6	Likely same (C)	Total for >20, total against =0, total for > last for + 10
7	Don't know	First, middle, last, and DoB all missing on one or both records, sex the same or missing on one or both records
7	Don't know diff sex	As don't know but without condition on sex
7	Don't know first partial agree	Middle, last and DoB all missing on one or both records, first names exactly the same to the length of the shorter string (e.g. Tom and Tomas)
7	Don't know last partial agree	As Don't know first partial agree but with condition on last
7	Likely different	Total for >50, total against <20
7	Probably different	Weak evidence against from first, last or DoB, total for > total against
8	Different – sub	Weak evidence against from up to two of first, last and DoB
9	Different other	Evidence against from first, last and DoB
7	Remaining	Any records not assigned to any of the above categories

These categories, the conditions and the distance scores were informed through clerical review of linking of the 2011 census and CCS datasets. Evidence of a match would be when the personal information (name, date of birth, sex and location) are the same or similar in the linked records. This evidence would be considered weaker if it still seemed possible that the two records represented the same person, but for this to be the case the information would need to have been recorded differently on the two sources, and/or that there were errors on one of both of the records.

Annex 5: Glossary

Term	Definition
Link	Two records that have been connected
Match	Two records that represent the same individual
Non-match	Two records that represent different individuals
False positive	A link that is a non-match
False negative	A match that has not been linked
Bad link	A link that has been clerically reviewed where the reviewer felt that it represented a non-match
Good Link	A link that either: • has been clerically reviewed where the reviewer felt that it represented a match, or • was used in 2011
DSE	Dual-System Estimation. A statistical process where two random selections are made of a population. By counting the number of individuals who appear in both selections, an estimate can be made for the total population.
NHSCR	National Health Service Central Register. An administrative dataset of patients registered with an NHS GP.

Annex 6: Information Governance

As with other linking to administrative datasets, this has been conducted in compliance with GDPR. The NHS Central Registrar was used as the administrative dataset for this quality assurance procedure, and the standard governance procedures were followed in this case. Only the Admin Data team will be working with this administrative data and it is only being used for quality-assurance processes.

More information on this can be found published on the website:

Data Protection Impact Assessment for use of NHSCR dataset
Quality Assurance report for use of NHSCR dataset for 2019