# **Methodology report on coverage matching for the 2021 Census**

**CRAG January 2019**

Matching, i.e. determining whether two (or more) different records belong to the same person (or household), plays a crucial part in the census statistical design and processing. It underpins the methodology for assessing coverage of the census, and therefore in producing high quality census based population estimates. This paper focuses on the three stages of census processing which include matching for the coverage assessment methodology.

This report is structured to talk about each of the three census matching processes under consideration (Resolve Multiple Responses, census to Census Coverage Survey, and census to census) in turn in each of the sections. In section 2, we briefly describe each of the three matches process, in section 3 we give a review of the lessons learned from 2011, and in section 4 we outline our intentions going forward. Finally, we discuss how the quality of the matching processes will be determined.

Linkage methods for other parts of the census processing, such as linking the census to administrative data, will be covered in a separate report.

## **Introduction**

Our strategy for developing matching methods for the 2021 Census is centred on achieving stringent quality targets (see section 2.2) concerning the small number of false positive (incorrect links) and false negative (missed links) matches that are allowed. These quality targets are required if the census is to meet its overall objective regarding the quality of the outputs. We will build on experience gained in 2011, but also incorporate methods that have been developed since then, both across the ONS and elsewhere. To achieve the quality targets, we will have to ensure an exceptionally low error rate is achieved in the automatic linkage process. This may mean that more clerical resolution is required than for other matching projects. Should it prove to take too long to achieve the required quality, we will advise Census and ask them what balance to strike between quality and timeliness.

The England/Wales and Scottish censuses differ from those of other countries in that we, as world leaders in this field, produce census estimates rather than census counts. Although other comparable countries, such as New Zealand, Australia, Canada and the United States do carry out a Post Enumeration Survey (PES), and use this to estimate the undercount and overcount in their census, they do not directly adjust their census counts [1]. This means that whilst the linkage requirements between the PES and the census in these countries are similar to those for the CCS to the census in England/Wales and Scotland, we will have to work to a much tighter time frame in order to produce census outputs in a timely manner.

## 2. The Three Census Coverage Matching Processes

### **2.1 Resolve multiple responses (RMR)**

This process is essentially an attempt to identify and resolve duplication in the census at the same geographical location. The process runs before any assessment of under or over-coverage. The idea is to resolve simple user duplication errors or duplications introduced by the census process itself. This process is carried out at small level geography (within household in 2011) to identify cases where people have responded to the census more than once. Examples of why such cases occur include respondents mistakenly filling in their own details on every page of the household form, or two members of the same household each completing the census form, perhaps one responding online and the other on paper. Records that are identified as within household multiple responses are removed at the RMR stage of processing. In 2011, approximately 237,200 person records were removed by the RMR process [2]. This process begins as soon as census returns start to arrive.

### **2.2 Census to CCS matching**

The Census Coverage Survey (CCS) is the key source of data for assessing under-coverage in the census. It is an independent enumeration of a sample of postcodes, designed specifically to measure coverage. A sample of postcodes are included in the CCS. In 2011, the CCS sampled 1.5 percent of all postcodes in England and Wales. The sample included 17,400 postcodes containing nearly 340,000 households [3]. Matching the census and CCS returns to a very high standard is critical to enable accurate estimation of the undercount in the census using a logistic regression model (although the underpinning model is based on capture-recapture methods) [4].

The matching requirements are to achieve a false positive rate of less than 0.1%, i.e. at least 99.9% of the matches we make must be correct; and a false negative rate of less than 0.25%, i.e. we must find at least 99.75% of all the matches. The denominator in both these calculations is therefore the total number of true positives – in 2011, this was determined by an extensive quality assurance process (see section 5).

These accuracy requirements are driven by the impact they have on the estimation process – they have a direct impact on the bias in the resulting population estimates (e.g. a 0.25% false negative rate results in a 0.25% overestimate of the population estimates, and given the census target is to achieve less than 0.5% bias this is hugely important). Census to CCS matching cannot be completed until we are in receipt of all the CCS and census returns for the CCS postcodes and surrounding areas.

### **2.3. Census to Census matching**

To enable measurement of the overcount in the census (after the RMR process above has been applied), we are also required to match the census to itself to find duplicate returns within a wider geography than that considered at the RMR stage. These cannot be automatically resolved by an algorithm (like in the RMR stage) as there is no information to say which of the records is in the ‘correct’ location. Thus, instead of resolving these duplicates, our approach is to measure the level of overcount, and then make a net under-coverage adjustment by subtracting the estimated over-coverage from the under-coverage. Examples of why such duplicates occur include students who are included on both their parents form at their home-address and at their term-time address, and children of separated parents who spend some time living at two different addresses. The overall level of overcount (of which duplication was one part, the other being people counted in the wrong place) in the 2011 Census was estimated to be 0.6 percent, or approximately 352,000 people [5].

This paper considers the lesson learned from 2011 and our intentions going forward to 2021 for each of the three matching processes described above.

## **Lessons learnt from 2011**

### **3.1 RMR**

In 2011, RMR used exact and rule based matching to find duplicate responses within households. Approximately 237,200 individuals were removed at the RMR stage of processing. Matches outside of household were not considered, although it was discovered during the census to CCS matching stage that there were numerous cases of duplicates occurring within the same postcode but with different address\_id codes. These cases of not-in-household duplicates were caused by errors in Address Base, and questionnaires being wrongly delivered to a neighbouring household, as well as genuine cases of people living at two different addresses such as children of separated parents.

The rule based matching relied on Soundex[[1]](#footnote-1) hash-codes of forenames and surnames matching. In addition, when resolving multiple individual responses, Levenshtein[[2]](#footnote-2) edit distance between the forenames was used.

To assess the effectiveness of the 2011 methods, a probabilistic method was used to find duplicates both within household and within postcode.

The probabilistic method found an additional 16,171 matches within household. Of these, over half (8,231) failed the Soundex test on forenames and just under a quarter (3,953) failed the Soundex test on surnames. Nearly 7,000 failed the Levenshtein test on forenames.

A clerical review of a small sample of 210 matches which were not found in 2011 and which failed the Soundex test on Forenames and/or Surnames and the Levenshtein test on Forenames was carried out to determine the type of mistakes that people commonly make when filling in the name field in the questionnaire. Common mistakes include: entering both first and middle name in the forename field, entering a title in the forename field and scanning errors (D 🡪 O, N🡪M, J🡪S etc). Some of these mistakes in the data can be corrected for. For example, using an ‘is contained within’ rule would correctly match names where a middle name is included in the forename variable such as DAVIDLEE and DAVID. This rule could also be used to correct for the inclusion of titles in the forename field, for example, MRKEITH and KEITH.

However, there were 167 (79.5%) good matches in the sample that failed the Soundex test on forenames, leading to the conclusion that using Soundex alone is not a reliable method of matching names in the RMR process. Experiments with alternative string comparison methods such as the New York State Identification and Intelligence System (NYSIIS) code did not improve the matching method. For example, forenames which include scanning errors such as MICAAHANTHONY and MICHAELANTHONY do not match on Soundex, Levenshtein or NYSIIS although they are the same person.

Not all the 210 records in the sample that were clerically reviewed were true matches. The sample included 31 record pairs that were probably false positives because they are twins. Although some of these could be automatically rejected because they were of different sexes or under the age of 30, we were still left with seven possible false positives. In all but one of these seven cases, the forenames contain a second name which is the same for both records, for example NORBERTSTEPHEN and ALBERTSTEPHEN which is why they have scored highly in the probabilistic matching. This will need to be accounted for in any probabilistic matching method that we propose for 2021.

The probabilistic method also found 44,972 matches that were within postcode but where the address\_id codes were different. Furthermore, in the ‘gold standard’ census to CCS matched records from 2011 there are 1,300 CCS IDs which link to two different census IDs with different address\_id codes. If the percentage of census duplicate records matched to a single CCS record is extrapolated to census records in areas not sampled by the CCS, we would have removed approximately 150,000 census duplicates of people in addition to the 237,200 removed by RMR according to Dini, Large, 2014 [2].

In 2021, we are therefore proposing to look for duplicates both within household and within postcode (or some other geography such as output area) at the RMR stage of processing. Further details about the overall 2021 proposed RMR approach are given in section 4.1. See [6] for further details of the comparison between the probabilistic method and the 2011 RMR results.

### **3.2 Census to CCS**

In 2011 the census to CCS matching was completed to the high standard required, achieving less than 0.1% false positives and less than 0.25% false negatives [7]. Approximately 650,000 people were included in the CCS data. Of these, around 450,000 (70%) were matched via automated methods including the use of exact matching and standard Fellegi-Sunter probabilistic. This left 30% for clerical resolution (deciding if a given pair of records are a match) and clerical searching (searching for a match when no possible matches are initially presented) [8]. See the automatic matching specifications report [9] for full details of the methodology used for the Census to CCS matching in 2011.

Although the quality of matching in 2011 was very good and will need to be equalled in 2021, we will need to perform the matching much more quickly in 2021. In 2011, matching the CCS to the census took the equivalent of 30 full time staff all working for 30 weeks. The matching was completed on 16th March 2012, with the last census questionnaires arriving in December 2011. This was achieved by employing both day- and evening- shifts of clerical matchers. To meet the 2021 Census objective of producing outputs within a year, the current assumptions in the processing timetable are that we will be able to complete the matching within an eight-week period in 2021.

There are several steps that we can take that will speed up the matching process without having a significant negative impact on the quality.

* Increase the number of matches made automatically – using matchkeys[[3]](#footnote-3) and probabilistic techniques, we can reduce the number of records that need clerical resolution to around 54,000 (9% instead of 30%) with only a minimal increase in the number of false positives so that the overall quality objectives are still achievable [10]. However, this is based on research using the 2011 Census and CCS data, and whilst this is the best data currently available, there is a risk that the 2021 data may be different (either worse or better).
* Decrease the need for clerical searching – using associative blocking and the Fellegi-Sunter algorithm [11,12] with low threshold we will be able to present clerical reviewers with candidate record pairs enabling them to find another 49,000 matches. Clerical matching in this way is much faster than clerical searching[[4]](#footnote-4) [10]
* Improve the clerical search system – it is inevitable that reviewers will have to do some clerical searching since not all records will have a matching record. The clerical search system used in 2011 was reportedly not as efficient or ergonomic as it could have been. It should therefore be possible to make improvements in this area [13].

Further details about the overall 2021 proposed Census to CCS approach are given in section 4.2.

### **3.3 Census to Census**

In 2011 a large-scale census to census matching exercise was carried out to identify duplicates in the census database. This focused heavily on those population groups that were likely to have been overcounted. The overall level of overcount that was estimated across England and Wales was approximately 352,000 people or 0.6% of the population [5].

Due to the high risk of false positives with a fully automated approach, the strategy was to search automatically and then sample the possible duplicate pairs for clerical review. Since the proportion, *P*, of census individuals who were counted more than once was expected to be small (*P<0.01*) and we needed to estimate with a good relative error, an inverse sampling technique was used in 2011 whereby records from each of 15 population groups were considered until 102 duplicates had been found in each group. The number 102 was chosen to give a coefficient of variance *CV(p)* of less than 10% [14].

The 15 population groups were, in priority order:

* Persons who have indicated they have a second residence on the census
* Students aged 18 to 25 by gender (2 groups)
* Armed forces personnel
* Children aged 0-4,5-15 (2 groups)
* Adults enumerated at a communal establishment aged 16-44, 45-74 and 75+ (3 groups)
* Individuals who complete the questionnaire using the internet aged 16-29, 30-49 and 50+ (3 groups)
* Everyone else by broad age groups 16-29, 30-49, 50+ (3 groups)

Each of the groups were sampled until 102 duplicate records had been found. The only exception to this was the first group where people indicated that they have a second residence and provided us a target postcode within which to search – this entire population group was sampled. The groups need to be non-overlapping, and the priority order dictates which takes precedent (for example, the sample from persons in a communal establishment will not include any students aged 18 to 25, armed forces personnel of children).

The duplicates found were all reviewed clerically to ensure that they were genuinely multiple returns from the same person rather than false positives.

The methodology for 2011 Census to Census matching was successful and it is proposed that a similar strategy be used in 2021.

However, there may be ways in which to improve the automatic matching so that not all duplicates need to be clerically reviewed. For example, in 2011 clerical reviewers checked that names were not really common (two different people called Robert is much more likely than two different people called Thunder), whether the occupation and marital status of the possible duplicates matched, and whether the other people in the household matched. For single person households, they also looked to see if the signatures were the same. At least some of these procedures could be automated in 2021.

Further details about the overall 2021 proposed census to census matching approach are given in section 4.3.

## **Intentions going forward**

### **4.1 RMR**

We will use a similar methodology to that developed for other census matching processes. This will be a combination of deterministic matchkeys followed by use of a probabilistic technique (Fellegi-Sunter and the Expectation Maximisation algorithm).

The Response Chasing Algorithm (RCA) needs an estimate of the response in each Lower Super Output Area (LSOA) on a daily basis once the census begins. For this reason, it will be necessary to run RMR frequently so that we can provide the RCA team with response rates adjusted for multiple responses. It is therefore proposed that the deterministic matchkey method is applied daily as this is expected to find a large proportion of the duplicate responses which, because they are exact matches or contain a previously agreed level of error, will not change status when the algorithm is re-run.

It is not feasible to run the probabilistic algorithm every day as it will require that we are in possession of a large proportion of the data. Furthermore, as more data becomes available, the results of the probabilistic algorithm could change with previous matches being declared non-matches and vice versa. We therefore propose that we run the probabilistic matching algorithm once only, when all the data are available. This will enable us to provide a final post-RMR stage dataset that can be used in the next stages of the processing stream.

To address the issue of not-within-household duplicates found at the census to CCS matching stage in 2011, we intend to widen the RMR search space from household to postcode (or some other area such as output area). Duplicates found within household at the RMR stage should be resolved at this stage. In addition, where entire households are found to be duplicates (and there are no extra people in either household), one of these can be removed at RMR stage. Rule based algorithms can be used to decide which person/household record to keep in these circumstances. For example, keeping the most complete response or the most recent response, or more likely, a slightly more complex hybrid approach as in 2011.

However, any duplicates that are found that are either:

* not within the same household
* not an entire duplicate household

should simply be flagged at the RMR stage, as the resolution algorithm for such cases would be difficult to code due to the potential number of permutations. These cases will therefore remain in the microdata and be in scope for the measurement of overcount. For census to CCS matching, they will be resolved as multiple matches if they are within a CCS postcode. The coverage adjustment process will then generalise the findings to the non-CCS areas.

### **4.2 Census to CCS**

#### 4.2.1 Deterministic Matching

Research is ongoing to develop matchkeys for automatic deterministic matching. At the current time, we have a series of 22 matchkeys which find 80.5% of the matches on the gold standard (the 2011 Census to CCS linked file).

Note that this figure of 80.5% should be treated with some caution for two reasons.

Firstly, the gold standard is not perfect. In some cases, links declared to be false links when comparing to the gold standard are true matches. Reasons for this include:

* The gold standard has linked the CCS record to a different census record which is a duplicate of the same person. These duplicates should be resolved during the RMR stage of processing in 2021.
* The CCS record is not on the gold standard at all (there are approximately 10,000 CCS records where this is the case – these cases cannot be explained).
* The CCS record was not successfully matched on the gold standard.

Secondly, we only know that some of the links we have found automatically are correct because they are on the gold standard. They would have been clerically reviewed in 2011 to have been confirmed as matches; images of the scanned questionnaires were used to assist this process. For example, cases where names do not match. We are currently reviewing the matchkeys to ensure that only definite matches are accepted automatically. Matchkeys where this is not the case can be used to generate pairs of possible matches for clerical review.

Table 1 gives details of the matchkeys that we have developed together with the number of one-to-one and one-to-many matches that they find. The sixth column gives the number of matches that do not agree with the gold standard, or where the CCS record is not included in the gold standard. A clerical review of these 525 records did not find any cases which look like incorrect matches. The final column gives the cumulative false negative rate when comparing the current matchkey method to the unique CCS IDs which have been matched on the gold standard (this denominator has been chosen so that we are comparing like with like).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1. Matchkeys developed for the census to CCS deterministic matching | | | | | | |
| **MK** | **Matchkey composition** | **Number of Matches** | | | **No. of one-to-one matches that are not on the GS** | **Cumulative False Negative Rate: one-to-one matches compared to unique CCS IDs *matched* on the GS** |
| **One-to-one matches** | **One-to- many matches** | **Total** |
| 1 | Fullname1, DOB2, Sex, Postcode | 245,296 | 1,480 | 246,776 | 73 | 63.6% |
| 2 | FNNPX3, Surname, DOB, Sex, UPRN4 | 28,643 | 115 | 28,758 | 4 | 59.4% |
| 3 | FNNPX, Surname, DOB, Sex, Postcode | 5,467 | 63 | 5,530 | 5 | 58.6% |
| 4 | Lev5(Fullname6) 2, DOB, Sex, Postcode | 119,764 | 1254 | 121,018 | 55 | 40.8% |
| 5 | Lev(Fullname)2, DOB, Sex, UPRN | 554 | 4 | 558 | 3 | 40.7% |
| 6 | Soundex(FNNPX), Soundex(Surname), DOB, Sex, Postcode | 8,675 | 94 | 8,769 | 2 | 39.4% |
| 7 | FNNPX, Surname, DOB, Postcode | 3,213 | 36 | 3,249 | 1 | 38.9% |
| 8 | StdLev6(FNNPX)>0.6, StdLev(Surname)>0.6, DOB, Sex, Postcode | 23,422 | 299 | 23,721 | 12 | 35.5% |
| 9 | StdLev(FNNPX, Surname)>0.4, (name transposed), DOB, Sex, Postcode | 1,606 | 33 | 1,639 | 5 | 35.2% |
| 10 | Forename27, Surname, DOB, Sex, Postcode | 126 | 1 | 127 | 0 | 35.2% |
| 11 | StdLev(Forename18)>0.4,  StdLev(Forename2)> 0.4, DOB, Sex, Postcode | 932 | 17 | 949 | 1 | 35.1% |
| 12 | Forename1(CCS)=Forename2(Cen) Surname, DOB, Sex, Postcode | 684 | 5 | 689 | 0 | 35.0% |
| 13 | Forename2(CCS)=Forename1(Cen), Surname, DOB, Sex, Postcode | 1,122 | 6 | 1,128 | 0 | 34.8% |
| 14 | FNNPX, DOB, Sex, UPRN | 17,146 | 119 | 17,265 | 17 | 32.3% |
| 15 | Surname, DOB, Sex, UPRN | 12,844 | 282 | 13,126 | 16 | 30.4% |
| 16 | FNNPX, Surname, DOB109, Sex, Postcode | 5,158 | 40 | 5,198 | 4 | 29.6% |
| 17 | FNNPX, Surname, Sex, DOB Year, Postcode | 18,537 | 142 | 18,679 | 10 | 26.8% |
| 18 | StdLev(FNNPX)>0.6, StdLev(Surname)> 0.6,  Lev(DOB) 1, Sex, UPRN | 10,841 | 117 | 10,958 | 10 | 25.2% |
| 19 | StdLev(FNNPX)> 0.6, StdLev(Surname)> 0.6, DOB matches or null, Sex, UPRN | 35,004 | 968 | 35,972 | 111 | 20.0% |
| 20 | StdLev(FNNPX)> 0.6, StdLev(Surname)> 0.6, DOB, Sex matches or null, UPRN | 1,245 | 22 | 1,267 | 3 | 19.9% |
| 21 | Fullname, DOB matches or null, Sex matches or null, UPRN | 782 | 14 | 796 | 164 | 19.7% |
| 22 | Fullname, DOB, Sex, Postcode Sector | 1,741 | 74 | 1,815 | 29 | 19.5% |
| Totals | | 542,802 | 5,185 | 547,987 | 525 | 674,146 |
| 1Fullname is a derived variable which is the concatenation of forename and surname  2DOB is full date of birth  3FNNPX is a derived variable (forename no prefix) which is forename with title MR, MRS, MISS removed  4UPRN is unique property reference number  5Lev is the Levenshtein Edit Distance  6StdLev is the standardised Levenshtein Edit Distance  7Forename2 is a derived variable containing all characters after the first space in forename  8Forename1 is a derived variable containing all characters before the first space in forename  9DOB10 – day of birth and month of birth match, years of birth are within 10 years of each other | | | | | | |

#### 4.2.2 Dealing with matches that are not one-to-one

The first three matchkeys allow for no error in the matching variables. After running all census and CCS records through these matchkeys, duplicate links (e.g. A-B and A-B) are removed. Conflict links (e.g. A-B and A-C) are flagged. For the exact matching it is assumed that these conflict cases occur when a single identity has been assigned two different unique IDs (either census or CCS).

All census and CCS records that were not matched exactly are then run through the remaining matchkeys. There are currently 19 extra matchkeys that allow for varying degrees of error. As before, duplicate record pairs are removed, and conflicts are flagged. The record pairs produced from matchkeys 4 to 22 are then concatenated together with the pairs from matchkeys 1 to 3, resulting in a final linked dataset.

The deterministic linkage used in matchkeys 4 to 22 uses a non-greedy approach. This means that if a record is linked in matchkey 4, it will still be considered for alternative links in matchkeys 5 to 22. Conflict links will inevitably occur because of this, however, as shown in Tables 2a and 2b, the size and number of clusters of links which contain conflicts is manageable. Clusters containing two IDs and one link are one-to-one links. We therefore plan to send any linking records in clusters of size 3 or more IDs, such as that shown in Figure 1, for clerical resolution.

Figure 1. Example of a cluster with four IDs, two Census (A & B) and two CCS (C & D), and three links made by matchkeys 4 and 10.

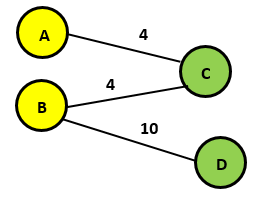


Table 2a shows the total number of IDs within a cluster, for all clusters produced in matchkeys 4 to 22. The maximum number of IDs is only seven showing these clusters do not get too big. Table 2b shows the total number of links within a cluster, again for all clusters produced in matchkeys 4 to 22. The maximum number of links within a cluster is 12 showing that these clusters do not get too complicated. Such clusters should therefore be easily resolved by a clerical matcher. Furthermore, the size and number of these clusters should be smaller in 2021 as improved RMR will remove more of the duplicate records.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 2a. Number of clusters with more than two IDs | |  | Table 2b. Number of clusters with more than 1 link | |
| **Number of IDs within a cluster** | **Number of Clusters** |  | **Number of links within a cluster** | **Number of Clusters** |
| 3 | 1,179 |  | 2 | 1,179 |
| 4 | 274 |  | 3 | 174 |
| 5 | 9 |  | 4 | 107 |
| 6 | 6 |  | 5 | 2 |
| 7 | 2 |  | 6 | 2 |
|  |  |  | 7 | 0 |
|  |  |  | 8 | 3 |
|  |  |  | 9 | 1 |
|  |  |  | 10 | 0 |
|  |  |  | 11 | 0 |
|  |  |  | 12 | 2 |

#### 4.2.3 Probabilistic Matching

We have also been designing blocking passes[[5]](#footnote-5) for use with the probabilistic method. As in 2011, this will make use of the Fellegi-Sunter and Expectation Maximisation (EM) algorithms [11,12] to match record pairs. It is expected that forename, surname, date of birth and postcode will be used as the variables for the expectation maximisation algorithm. An examination of record pairs on the gold standard which have not been matched by the matchkeys shows that approximately one third of them agree on two of fullname, postcode and date of birth. It is therefore expected that these will be matched by the probabilistic algorithm.

The blocking passes shown in Table 3 were used to create a suitable set of candidate pairs and to allow error in a variety of variables, which enabled identification of as many matches as possible. These blocking passes created a total of 10,909,950 candidate pairs.

|  |  |  |
| --- | --- | --- |
| Table 3. Blocking passes used for census to CCS probabilistic matching algorithm, conditions and number of pairs generated (after deduplication). | | |
| **Blocking Pass** | **Conditions** | **Number of Pairs** |
| 01 | UPRN | 1,685,121 |
| 02 | DOB, Postcode Sector | 729,418 |
| 03 | Lev1(Fullname) <3, Day, Month, Sex | 1,432,513 |
| 04 | Lev(Fullname) <3, Year, Month, Sex | 1,058,346 |
| 05 | Lev(Fullname) <3, Day, Year, Sex | 652,470 |
| 06 | StdLev2(Forename) >0.4, StdLev(Surname) >0.4, DOB, Sex | 1,119,880 |
| 07 | StdLev(Forename) >0.6, StdLev(Surname) >0.6, Day, Year, Sex | 716,057 |
| 08 | StdLev(Forename) >0.6, StdLev(Surname) >0.6, Day, Month, Sex | 1,562,177 |
| 09 | StdLev(Forename) >0.6, StdLev(Surname) >0.6, Year, Month, Sex | 1,151,340 |
| 10 | Lev(Fullname) <2, Postcode Sector, Sex | 427,279 |
| 11 | Lev(Forename) <2, Lev(Surname) <2, DOB, Postcode Sector | 375,349 |
| 1Lev is the Levenshtein Edit Distance  2StdLev is the standardised Levenshtein Edit Distance  (see footnote on page 3 for an explanation of how these are calculated) | | |

The above candidate pairs were input to the EM algorithm and the and values[[6]](#footnote-6) were calculated as shown in Table 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 4. - and -values calculated using the EM algorithm | | | | |
|  | **Forename** | **Surname** | **DOB** | **Postcode** |
| **-value** | 0.866 | 0.811 | 0.996 | 0.998 |
| **-value** | 0.524 | 0.337 | 0.222 | 0.257 |

After calculating the - and -values and associated agreement and disagreement weights, scores are calculated for each candidate pair. A threshold is selected and only pairs scoring above this threshold are accepted as matching record pairs. Records that had two or more candidate pairs with the same score were removed and added to a separate table to be sent to clerical review (if their score is above the threshold). This resulted in a set of 669,764 unique pairs that were compared to the matches created using the deterministic matchkey model described in section 4.2.1 Deterministic Matching.

#### 4.2.4 Comparison of deterministic and probabilistic methods

Of the 547,987 CCS records that were matched using the deterministic model, 542,764 (99%) were paired with the same record using the probabilistic method. 3,799 of the CCS IDs did not appear in the automatically accepted probabilistic file because they had multiple candidate pairs with the same score and were therefore in the file for clerical resolution. There were also 1,424 records that the probabilistic method paired with a different census record to the deterministic method. From a clerical review of these 1,424 records, it appears that in many cases the different census records are duplicates. The deterministic method has linked to one of the records and the probabilistic method has linked to another version of the same record. It is hoped that improving the RMR process in 2021 will reduce the number of conflicts caused by duplicates.

There are some examples of matching records where the forename and surname have been reversed from the census and CCS. The deterministic method uses a matchkey to account for this whereas the probabilistic method will not consider this type of data error and so may offer a pair with a higher score which is not the correct pair thereby generating a false positive match. On the other hand, the probabilistic method made matches where there was agreement on date of birth and postcode but a lot of error in name, and the deterministic method failed to link these records.

Although both methods have made many of the same links, examples can be seen where one method is matching individuals that the other method has missed. This suggests that there is strength in developing a matching method that involves both deterministic and probabilistic matching as this will give the best chance of making the highest possible number of correct links.

Further work is required to assess the candidate pairs made using the probabilistic method but not the deterministic method, to determine how many of these could be accepted automatically, and also to assess where the threshold should be set. The candidate pairs will also be compared to the gold standard to assess the precision (the proportion of links made that are true matches) and recall (the proportion of all true matches that have been found).

The threshold for automatically accepting pairs generated by the probabilistic algorithm will be set high to minimise the risk of false positives. However, a lower threshold will be used to generate pairs that can be clerically reviewed.

It is hoped that as much as 91% of the census to CCS matching can be done using this combination of deterministic and probabilistic automatic methods. However, we need to treat this figure of 91% with some caution as it may be over-optimistic for several reasons.

* The matchkeys can only be fitted a priori to the 2021 data, and we must make assumptions as to what that data will look like. Overfitting the 2021 model to the 2011 data could cause problems if the 2021 data is significantly different to the 2011 data. Since the 2021 data is to be collected primarily online rather than on paper questionnaires, it is reasonable to assume that the 2021 data will indeed be significantly different to that collected in 2011. We can mitigate against this risk by using a probabilistic method in combination with the matchkeys. Since the EM algorithm does not require a training set to generate the weights it is better able to cope with data that is not exactly as expected compared to the deterministic method.
* If response rates to either the 2021 Census, or the CCS, or both, are lower than those achieved in 2011 (94% and 90.8% respectively) then the pool of residuals (unmatched records) will be larger than in 2011. This will require extra clerical work as it is not possible to declare these records as unmatched until a search for a matching record has been conducted.

#### 4.2.5 Clerical Matching

Increasing the number of automatic matches made, from around 70% of person records in 2011 to a projected 91% of person records in 2021, will speed up the matching process. However, there will still be the need to clerically review:

* record clusters generated by the deterministic matching where there is a one-to-many match;
* record pairs where there is more than one match with the same probabilistic score;
* residual records that have not been matched by either the deterministic or probabilistic methods.

Clerical matching is much faster when reviewers are presented with candidate pairs of records rather than having to search for a match themselves. We are therefore devising methods of generating candidate pairs for as many of the residuals as possible. These methods will include using the probabilistic method with a low threshold, and using matchkeys not deemed accurate enough for automatic matching.

#### 4.2.6 Household Matching

There is also the need to match households. In 2011, households were matched using a mixture of deterministic and probabilistic algorithms. Postcode, address, tenure, head of household surname, accommodation type, resident count, telephone and number of rooms, were used as matching and/or blocking variables. In 2021, we could use a similar method, potentially also including unique property reference number (UPRN).

An innovative method that matches households by creating a hash key using the sums of the components of the dates of births and sexes of the household members has been proposed. The benefit of this method is it that is does not rely on names being correct. However, it does rely on having all the same people in the household on the census and CCS, and it has not yet been proven to be more effective in terms of quality or number of matches made, compared to using UPRN or raw address data. Further research is required to determine the best method for matching households and for assigning individual and extension forms to existing households.

In 2011, households were matched before people. We are currently considering whether this is the correct approach for 2021, or whether it is more effective to match people and households independently since household matching can be thought of as matching household spaces rather than people in houses. Note that in using the second approach we would still make use of attributes of people in order to make connections between households. Further research is required in terms of the exact requirements of the household matching with regards to what is necessary for the estimation processes.

### **4.3 Census to Census**

Work is ongoing to review the methods of 2011 and program these in PySpark so that the power of the distributing access system (DAP) can be harnessed. It is likely that the 15 groups of people who are most likely to be overcounted (see section 3.3 Census to Census) will remain the same in 2021. The only exception is the group of ‘individuals who complete the questionnaire using the internet’. In 2011, 16% of households chose to respond to the census online. However, planning for the 2021 Census is based on an estimated 75 percent online completion rate. It therefore makes sense to change this group to be the ‘individuals who complete the questionnaire on paper’.

Instead of sampling within the 15 groups until a sufficient number of duplicates are found for estimation purposes, it might be feasible to use the 15 groups as blocking categories and find all the duplicates within each group using a probabilistic method. This method is currently being tested. A sample of the matches found using this method could be clerically reviewed to assess their quality. Further work is also required to assess the possibility of automating some of the checks made by reviewers in 2011 so as to lessen the burden of clerical review and speed up the matching process.

## **Measuring matching quality**

Methods are quality assured as they are developed, for example by comparing with the results of the 2011 matching and by the use of clerical review. However, there will be the need to quality assure our matching processes during the live census processing.

For RMR this may take the form of reviewing the number of households and individuals removed by the process to ensure that it is within expected bounds (for example compared with the 2011 RMR process), and some clerical review of the identified duplicates.

In 2011, there was an extensive quality review of the census to CCS matching to ensure that the levels of error in the matching process were within the quality targets. This was done after the matching had been completed. This was achieved by re-matching 1,084 CCS postcodes (6.2%). The re-matching focused on the hardest to count areas, which tended to have the lowest automatic match rates and so the largest workloads for the clerical matchers. The sample contained 83,356 original matching decisions (approx. 17% of the original matching decisions). The results of this evaluation showed that the matching quality was very high overall, and also in each hard-to-count area. It also showed that there was consistency between the results of the day- and evening-shift matchers. See [7] for further details.

It is worth noting that in theory an estimation method that takes into account the linkage errors may be adopted [15,16]. Such methods require estimation of linkage error rates from a rematch study which is similar to a study aiming to assess the quality of the linkage process. There may be at least two strategies employed in the case of utilising one of the linkage error corrected estimation approaches. The first strategy aims to achieve very high-quality linkage in the first place and then is followed by the rematch study. The rematch study estimates the linkage error based on a sample of record pairs extensively reviewed by the most experienced clerical staff. The estimated linkage errors are then fed into a chosen linkage error correcting estimation approach to increase the quality of estimation. The second strategy somehow relaxes the quality of the linkage in the first place and assumes that the rematch study is capable of providing high quality estimates of the linkage errors and the linkage error corrected population size estimates.

Despite a recent increase of interest in the field of linkage error corrected population size estimation, the majority of achievements are mainly of theoretical interest and do not span all the complexities of the census coverage assessment project. Most of the estimators proposed so far perform in a relatively simple set up [15,17] and a huge effort is required to adopt these methods to achieve the census quality targets. In addition, using the second strategy (lower linkage quality in the first place, then high quality rematch study) puts the census quality targets into risk as some recent work shows that it could lead to a substantial variability in the linkage error corrected population size estimates when the linkage error rates are high.

Given that the linkage error correct population size estimation is not currently mature enough to meet the census quality targets, it is recommended to put substantial resources into achieving the highest quality linkage.

## **Conclusion**

We have methods and processes in place that will enable us to complete the census coverage matching processes to the high standard required. However, there are factors, outside of our control, that will determine the time needed.

In particular, if response rates for the census and/or CCS are lower than in 2011 this will have a huge impact on the amount of clerical resolution and searching that will be necessary for the census to CCS matching.

Furthermore, deterministic methods are being optimised based on 2011 data. Whilst this is the best data currently available for testing our methods, it is probable that the 2021 data will be significantly different. For this reason, we need to also include probabilistic linking methods. This is the best way of having an overall method that is sufficiently flexible to deal with new data, which does not necessarily fit our rules, without the need for a training set or prior model specification.

We are confident that we will be able to complete the census coverage matching to the required high quality. The challenge is to automate the processes sufficiently, without sacrificing quality, so that we can deliver within the required timelines.

## **References**

[1] Statistics New Zealand (2013) Coverage in the 2013 Census based on the New Zealand 2013 Post-enumeration Survey. Available at <http://archive.stats.govt.nz/browse_for_stats/population/census_counts/report-on-2013-post-enumeration-survey.aspx>

[2] Dini, E., Large, A. (2014) *Overcount in the 2011 Census: an Assessment of the Methodology* in Survey Methodology Bulletin No. 72, Spring 2014

[3] Office for National Statistics (2012) 2011 Census Coverage Survey Summary. Available at <https://webarchive.nationalarchives.gov.uk/20140113111903/http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-censu++s-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/census-coverage-survey--ccs-/index.html?translation-component=&calling-id=77-6743-4&currLang=English&format=contrast>

[4] Racinskij,V. (2018) Coverage Estimation Strategy for the 2021 Census of England & Wales. Unpublished Internal Report. Available on request.

[5] 2011 Census: Methods and Quality Report, *Overcount Estimation and Adjustment*, July 2012

[6] Shipsey, R. (2018) [Methodology report on Resolve Multiple Responses](https://share.sp.ons.statistics.gov.uk/sites/MTH/Cen/2021/Data_Linkage_Methodology/RMR/RMR_scoping_report_2021.docx). Unpublished Internal Report. Available on request.

[7] Office for National Statistics (2013). An Assessment of the Quality of the Matching between the 2011 Census and Census Coverage Survey. Available at: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/census-coverage-survey--ccs-/matching-quality-report-ns.pdf>

[8] Office for National Statistics (2012). Automatic Match Rates for the 2011 Census to the Census Coverage Survey. Available at: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/automatic-match-rates-for-the-2011-census-to-the-census-coverage-survey.pdf>

[9] Baillie,M., Large, A. (2010) Automatic matching specifications report. Unpublished Internal Report. Available on request.

[10] Cummins, S., Jones, P., Gammon, S. (2016). Improvements in methodology for matching the 2021 Census to the Census Coverage Survey. Proceedings of the 21st GSS Methodology Symposium. London, 2016. Available from <https://gss.civilservice.gov.uk/events/21st-gss-methodology-symposium/>

[11] Fellegi, I., Sunter, A. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, Vol. 64, No. 328.

[12] Winkler, W. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 667-671.

[13] Shipsey, R. (2018) [Clerical searching for the 2021 Census to CCS matching](https://share.sp.ons.statistics.gov.uk/sites/MTH/Cen/2021/Data_Linkage_Methodology/Census_CCS/Clerical_searching_presentation.pptx). Unpublished Internal Presentation. Available on request.

[14] Abbott, O., Large, A. (2009) Measuring the level of duplicates in the 2011 Census. Available at <https://www.researchgate.net/publication/305710145_Measuring_the_level_of_duplicates_in_the_2011_Census>

[15] Ding, Y., Fienberg, S. (1994) Dual System Estimation of Census Undercount in the Presence of Matching Error. Survey Methodology, December 1994. Volume 20 pp 149-168. Statistics Canada.

[16] De Wolf, P., Der Laan, J., Zult, D. (2018) Connecting correction methods for linkage error in capture-recapture. Available at <https://www.cbs.nl/en-gb/background/2018/18/connecting-correction-methods-for-linkage-error-in-crc>

[17] Chipperfield,O., Bishop, G., Campbell, P. (2011) Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. Available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11444-eng.pdf?st=HdSWzR5w>

1. Soundex is a phonetic coding system that produces a 4-character hash-code of the name. It does not cope well with very short names, names of an Oriental origin, or names which have a high percentage of vowels. Furthermore, because it is designed to correct for phonetic mistakes e.g. B, P, F and V are all treated as the same character because they sound similar, it does not appear to work so well with the types of errors that occur in written names, such as D being mis-scanned as O or IVI begin mis-scanned as M. [↑](#footnote-ref-1)
2. Levenshtein edit distance measures the number of changes required to change one name into another. For example, TIM and TOM have a Levenshtein edit distance of 1, and SAMANTHA and SAMATHA also have a Levenshtein edit distance of 1. To prevent bias due to length of name (a long name where there is one mistake is more likely to be correct than a short name where there is one mistake) the following formula for standardised Levenshtein was used to declare a match.

   Thus TIM and TOM score 0.67 and fail this test, but SAMANTHA and SAMATHA score 0.875 and pass. [↑](#footnote-ref-2)
3. A matchkey is a combination of variables on which two records must agree in order to be declared a match. See Table 1 in section 3.2 for examples of matchkeys. [↑](#footnote-ref-3)
4. Based on 2011 figures, a clerical reviewer could review approximately 40 record pairs per hour when presented with candidate pairs (clerical matching). However, they could only review 4-6 records per hour if they had to search for a match manually (clerical searching). [↑](#footnote-ref-4)
5. Blocking passes are used to bring together record pairs that have something in common so that they can be compared with one another and assigned a score. The higher the score, the more likely that the pair are a match. [↑](#footnote-ref-5)
6. The -value is the probability that the variable agrees given that the pair is a true match. Thus the -value is a measure of the data quality. The -value is the probability that the variable agrees given that the pair is not a true match. Thus the -value is the likelihood of matching by chance. [↑](#footnote-ref-6)