Census to census matching strategy 2021

Since being taken to the Methodological Assurance Review Panel (MARP), work in the area covered by this paper has progressed and has now been completed. This version reflects these changes, as well as incorporating panel comments.

Executive Summary

We will match the 2021 Census dataset to itself in order to find duplicate responses, i.e. individuals who respond more than once. This will enable the overcount in the census to be estimated. We propose to use an inverse sampling method broadly similar to that used in 2011. However, we will also implement an automated checking algorithm (ACA) that will automatically accept or automatically reject clear cut cases without the need for any clerical resolution. We expect this algorithm to reduce the amount of clerical resolution required for census to census matching by around 70% without compromising on the accuracy of the matching.

Introduction

Census to census matching is required to estimate the overcount – that is, people who have been enumerated in the census more than once. Although overcount in the census is less of a problem than undercount, it is too large a problem to ignore. In 2011 it was estimated that there was an overcount of 0.6% or around 352,000 people. Census to census matching for overcount was first introduced to the census estimation processes in 2011 due to the change to post-out rather than hand delivery of census questionnaires. Evidence from other countries such as Australia, who run a similar census operation to England and Wales, suggests that the level of overcount in 2021 is likely to be higher than in 2011 [1].

The difference between census to census matching and resolve multiple responses (RMR)

The aim of census to census matching is to find people who have been included on more than one census return (duplicates) at different locations. Therefore, census to census matching does not rely on geographical or address matching, although matching second or alternative addresses can be used to confirm duplicates. Duplicates found during census to census matching cannot be resolved as there is generally no way of knowing which return is the correct one, and in some cases, both are correct. Census to census matching is one of the later stages of census processing. Examples of why this type of duplication occurs include:

- People who have two or more houses, who complete a census return for each address
- People who move to a new house during the census period and complete a census return at both their old and new addresses
- People who live and work in two separate locations
- Students who are fully enumerated at both their term time address and their family home
- Children of separated parents who spend some time living with each parent
- People who go into a hospital or care home and who are included on both a communal establishment form and their own household form.

In contrast, the Resolve Multiple Responses (RMR) process, which is described in [2], aims to find and resolve duplicates that occur at the same location. Duplicates found during RMR can be resolved by merging them into a single return. RMR is one of the first stages of the census processing pipeline. This type of duplication can occur because:

- Someone has filled in their own details more than once on the same form
- Two people at the same address have each completed the census, maybe one on paper and the other online
- Errors in the collection process have led to two or more census forms being delivered to the same household.

2011 Census to census matching process

In 2011, an inverse sampling method was used to estimate the prevalence of overcount in different groups and regions. 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 [3] whereby records from each of 15 population groups in each of ten regions were considered until 102 or more duplicates had been found in each group (150 groups altogether). The number 102 was chosen to give a coefficient of variance $CV(p)$ of less than 10%. The 15 population groups were defined using the 2001 Longitudinal Study and 2001 Census [4].

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)

The 10 regions were:

- North East England
- North West England
- Yorkshire and Humberside
- East Midlands

- West Midlands
- East of England
- London
- South East England
- South West England
- Wales

Each of the groups above were sampled in each of the ten regions until 102 or more 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 were non-overlapping, and the priority order dictated which took precedence. The duplicates found were all reviewed clerically to ensure that they were genuinely multiple returns from the same person rather than two different people with similar names.

1. Research for the 2021 Census to census matching process

5.1. Can we match the whole of the census to itself?

Ideally, the estimation team would like us to be able to say, for each person in the census, whether they are a duplicate or not. Due to the increase in computing power since 2011 and the ability to utilise distributed computing and parallelise processes in the Data Access Platform (DAP), we have considered running a probabilistic matching method for all of England and Wales.

Using the 15 groups above as blocking passes, together with a fuzzy match on name[1] and exact match on date of birth and sex, the probabilistic algorithm generates 38,809,506 candidate pairs in approximately 39 hours. Using clerical review to set a threshold below which we reject all candidate pairs, leaves us with 930,353 candidate duplicate pairs. Therefore, whilst it is computationally feasible to match census to census for all of England and Wales, clerically reviewing all the candidate pairs to ensure that they are genuine duplicates is not feasible within the time and cost constraints of the census.

Can we reduce the amount of clerical review needed?

In 2011, clerical reviewers considered all the candidate duplicate pairs and then decided if they were genuine duplicates or not. For example, if there are two people called JOHN SMITH both born on the same day one of whom is a doctor and the other is a carpenter, then these are most likely two different people. However, if there are two people called PERSEPHONE ASQUITH both born on the same day one of whom is a teacher and the other is a lecturer, then these are likely to be duplicates of the same person.

1 Two names are considered to be a fuzzy match if the standardised Levenshtein edit distance between them is greater than 0.6 meaning that at least 60% of the characters match.

We have tried to replicate automatically the reasons for saying a candidate pair is or is not a genuine duplicate in clear cut cases. This has enabled us to split the candidate pairs up into three sets:

- Accept automatically as a duplicate
- Reject automatically as a duplicate
- Send to clerical review

This process is described in the next section.

Automated Checking Algorithm

Candidate duplicates who match[2] on all key matching variables (forename, middle name, surname, date of birth and sex) and either match or have missing values for all secondary matching variables (ethnicity, occupation, marital status and country of birth) are checked to see if the spouse is different. If the spouse is clearly different, the candidate pair is rejected. Otherwise the pair is sent through a series of confirmation steps as shown in table 1. If the candidate duplicate pair passes one of the confirmation steps, it is automatically accepted. Candidate duplicate pairs who have not passed a confirmation step, but who have an uncommon name (see section 6.2) are sent for clerical review, those with a very common name are automatically rejected.

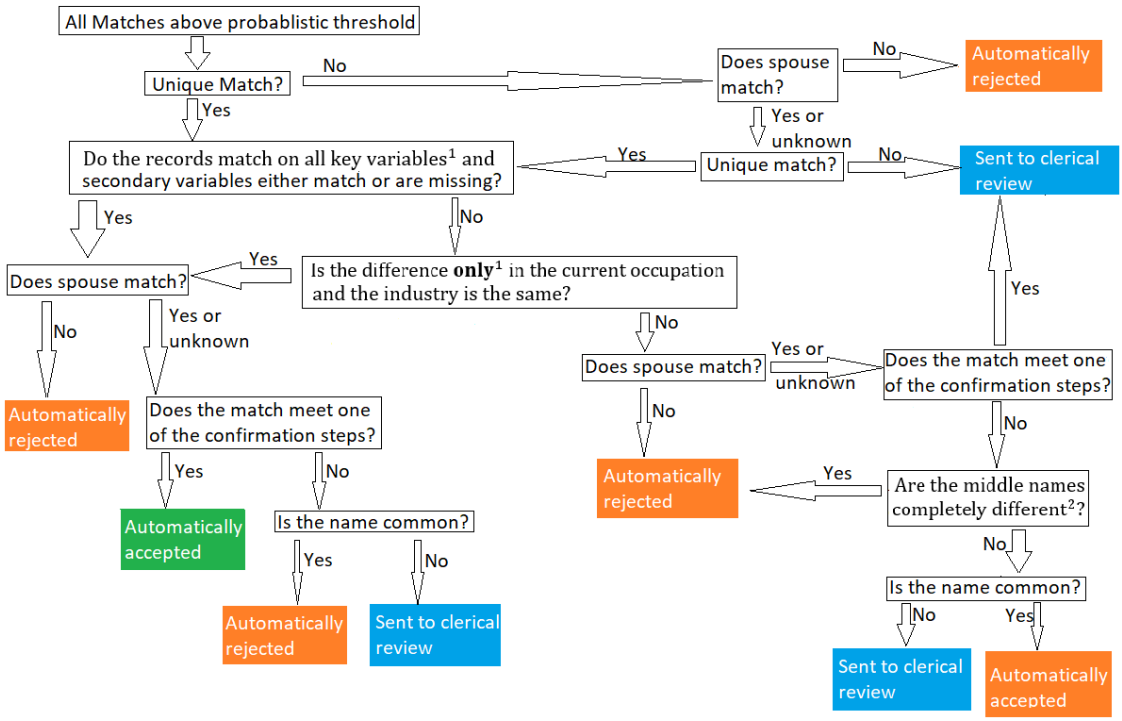
The remaining candidate duplicates must have differences in the secondary variables or have names with a standardised Levenshtein edit distance of less than 0.8. If the only difference is the occupation, but the industry codes are the same, then the candidate duplicate is sent through the same confirmation process as described above. In all other cases, the candidate still goes through the rejection/confirmation process as above, but is either sent for clerical review or rejected, never automatically accepted. The threshold for what constitutes a common name is lower here. Candidate pairs that do not meet a confirmation step and which have different middle names (i.e. they do not meet any of the matching criteria for middle names as described in footnote 2) are automatically rejected.

Note that only unique candidate pairs are sent through this process. Where we have a cluster of three or more possible duplicates, we first check to see if any have different spouses. If this is the case, we remove this candidate or candidates from the cluster. If the cluster is now comprised of just a pair of records it is sent through the automated checking algorithm. Otherwise the cluster is sent for clerical review.

---

2 We consider names to match if they score more than 0.8 using the standardised Levenshtein edit distance function. This is to allow for scanning and spelling errors. In addition, if one of the middle names is an initial only, and this matches the first letter of the other middle name, then this is considered to be a match. If one or both of the records does not have a middle name this is not penalised i.e. the names are still considered to be matching if the forename and surname match. Research on middle names was carried out using the 2011 Patient Register data since middle names were not collected on the 2011 Census.

Figure 1. shows how a candidate duplicate pair may be automatically accepted, automatically rejected or sent for clerical review.



1 A match on name here means sLev(forename, middle name, surname) >0.8, with the exception that middle names can be missing on one or both records, or if only the initial is given on one of the records, it matches the first letter of the middle name on the other record.

2 Completely different middle names mean that these middle names have a sLev(middle name) <0.8, the initials do not match and neither middle name is missing

Figure 1. Automated Checking Algorithm

Table 1 describes the confirmation steps and rejection steps that are used to automatically accept or reject a candidate pair of duplicates.

Table 1. Confirmation and rejection steps

Confirmation Step	Description
1	There are other duplicates within the two households of the candidate duplicate – finding a whole family where everyone is duplicated is evidence that this is a duplicate return, possibly the family moved to a new house during the census period and completed a form at both locations.
2	The candidate duplicate is a student and in one household appears to be living in their family home and in the other is in student accommodation or a household made up of students.

3	The candidate duplicate is included in a care home or hospital and on a household form in a nearby location.
4	On one (or both) returns the candidate duplicate gave a second address and this is the location where we have found the other response.
5	The address of the candidate duplicate matches the address one year ago of the other response.
6	The workplace address of the candidate duplicate matches on both responses.

Rejection Step	**Description**
1.	The candidate duplicate has a different spouse (the standardised Levenshtein score of the forename and surname are both less than 0.5, or more than half the digits in the date of birth are different)
2.	The candidate duplicates do not meet any of the confirmation steps and have middle names which are completely different (the standardised Levenshtein score is less than 0.8, initials do not match and neither middle name is missing)
3.	The candidate duplicate has matching key and secondary variables, but does not meet any of the confirmation steps and has a very common name i.e. the forename and surname combination appears in the census data in the year of birth of the candidate pair 42 or more times.
4.	The candidate duplicates do not meet any of the confirmation steps, have differences in the secondary variables and have a common name i.e. the forename and surname combination appears in the census data in the year of birth of the candidate pair 23 or more times.

Using commonness/rarity of names

For those candidate pairs who have not met any of the confirmation criteria in Table 1, we next consider how common the name is. According to the birthday paradox [5], if there are just 23 people in a room, the chance that two of them share a birthday is greater than 50%. By the same logic, if there are 23 people with the same name (forename and surname together), born in the same year, then the chance that two of them share a birthday is greater than 50%. So, for people with *common names* who do not meet any of the confirmation criteria, we assume that these are two different people who happen to have the same name and birthday, and automatically reject the match.

The thresholds for determining whether or not a name is common are as follows:

- In cases where all key and secondary variables match, research has shown that in cases where the names are very common the match was rejected by the clerical reviewer in 2011. Hence, the threshold for a common name in these cases is a name (forename and surname together) which occurs 42 or more times in the year of birth. At this threshold there is a 90% chance of two people with the same name and birth year having the same birthday. Hence, candidate pairs who seem to be very good matches will not be automatically rejected unless the name is very common.

- In all other cases, i.e. there is some disagreement or missingness in the secondary variables, the threshold for a common name is set at 23 or more in one year of birth. Hence, candidate pairs who have the same name, which is quite common, and who have not met a confirmation step, and also have some differences in the secondary variables will not be sent to clerical review but are automatically rejected.

Note that in cases where names do not match exactly, for example due to scanning errors, it is the count of the most common name that is used. For example, the name pair (ROBERT_SMITH, ROBERT_SNITH) is treated as common even though the second name in this pair does not occur many times.

Automated checking algorithm counts of outcomes

After a clerical review of decisions made by the algorithm, the probabilistic algorithm was changed to include an additional blocking pass – people who said that they had a different address one year ago, and the threshold was raised. This led to 796,591 candidate pairs being put through the automated checking algorithm (ACA). Table 2 shows the number of candidate pairs which ended up in each set (accept automatically, reject automatically or send to clerical review) and the stage at which they were confirmed or rejected.

Table 2. Outcome of the ACA with the number and percentage of candidate pairs at each stage.

Outcome	Stage	Number of candidate pairs	Percentage of candidate pairs
Confirm automatically	**Total**	**455,889**	**57.23**
	Exact matches & matches with one or more missing secondary variables matches & matches with different occupation but same industry, meets a confirmation step:		
	1 (other duplicates in HH)	83,579	10.49
	2 (student)	140,528	17.64
	3 (care home/hospital)	1,338	0.17
	4 (second address)	214,934	26.98
	5 (address 1 year ago)	11,787	1.48
	6 (workplace address)	3,723	0.47
Send to Clerical Review	**Total**	**213,770**	**26.84**
	Match is not unique	51,755	6.5
	Exact matches & matches with one or more missing secondary variable & matches with different occupation but same	47,303	5.94

	industry, does not meet a confirmation step, name is not common		
	One or more secondary variables different, meets a confirmation step:		
	1 (other duplicates in HH)	6,213	0.78
	2 (student)	13,770	1.73
	3 (care home/hospital)	173	0.02
	4 (second address)	17,024	2.14
	5 (address 1 year ago)	2,928	0.37
	6 (workplace address)	775	0.1
	One or more secondary variables different, does not meet a confirmation step, name is not common	73,829	9.27
Reject Automatically	**Total**	**126,932**	**15.93**
	Non unique matches with different spouses	8,976	1.13
	Exact matches with different spouses	2,740	0.34
	Matches, with missingness with a different spouse	3,070	0.39
	Matches, with only current occupation different but in same industry with a different spouse	1,583	0.2
	Matches with one or more different secondary variables, with a different spouse	53,094	6.67
	Exact match, does not meet a confirmation step, name is common	169	0.02
	Matches with missingness, does not meet a confirmation step, name is common	7,361	0.92
	Matches with a different occupation and the same industry, does not meet a confirmation step, name is common	347	0.04
	Matches with one or more different secondary variables, does not meet a confirmation step, name is common	49,592	6.23

We can see from Table 2 that the number of candidate pairs sent to clerical review is reduced to 213,770 which is 27% of the original number of candidate pairs. Using this algorithm could therefore greatly reduce the amount of clerical effort required.

Table 3 shows the number of candidate pairs from each overcount-group that were automatically accepted, rejected or sent to clerical review. These figures are based on 2011 data and using the entire population of England and Wales rather than the inverse sampling method.

Table 3. Number of candidate duplicate pairs that are automatically accepted, rejected or sent to clerical review from each group

Group	Total number of possible matches	Number automatically accepted	Number sent to clerical review	Number automatically rejected
Persons with a second residence	453,461	397,570	54,302	1,589
Male students aged 18-25	9,091	3,406	5,346	339
Female students aged 18-25	8,780	3,714	4,887	179
Different address one-year ago	17,930	13,943	3,734	253
Armed forces personnel	1,075	164	884	27
Children aged 0-4	10,480	3,816	5,743	921
Children aged 5-15	29,482	7,075	18,442	3,965
Adults in a CE aged 16-44	6,906	931	5,474	501
Adults in a CE aged 45-74	2,519	601	1,404	514
Adults in a CE aged 75+	2,132	175	1,852	105
Responded online aged 16-29	21,132	4,007	12,997	4,128
Responded online aged 30-49	40,710	4,527	16,602	19,581
Responded online aged 50+	28,299	2,039	9,702	16,558
Everyone else aged 16-29	26,831	3,697	17,152	5,982
Everyone else aged 30-49	54,893	4,537	23,772	26,584
Everyone else aged 50+	82,870	5,687	31,477	45,706
Total	796,591	455,889	213,770	126,932

How well does the automated checking algorithm work?

Clerical review of the decisions made by the algorithm have been ongoing during its development. As a result, we have made improvements on each iteration including:

- Previously, candidate duplicates who matched exactly on all key and secondary variables were accepted automatically without any further confirmation. This caused too many false positives. Now all candidates are put through the confirmation steps before being automatically accepted.
- Previously, having an uncommon name was considered to be a confirmation step i.e. if a candidate duplicate pair matched on all key and secondary variables and had an uncommon name then the pair was accepted automatically. This was found to cause false positive matches. A candidate pair now has to meet one of the other confirmation steps in order to be automatically accepted.
- Too many candidates were going to clerical review because they had missingness in the secondary variables. We now accept automatic matches with missingness (not differences) in the secondary variables if a candidate passes a confirmation step.
- Candidates were going to clerical review and then rejected because they had different spouses – this has now been automated so that more candidates are automatically rejected. We previously tried to make use of other relationships in the relationship table, but this caused too many false negatives.
- If occupation is different, but industry is the same, then candidates are put through the confirmation steps and could be accepted automatically.
- Use of middle name to help reject candidates where middle name is different.

A clerical review of 1,496 automatically accepted pairs found that 123 (8.2%) were false positives. These were all caused by automatically accepting pairs with an uncommon name when no other confirmation criteria had been met. This step was removed, and a subsequent clerical review found no false positive matches had been automatically accepted. In addition, a clerical review of 1,267 automatically rejected pairs found that there were no false negatives.

Proposed method for 2021 Census to Census matching

Although the ACA can be used to cut down the amount of clerical review needed for census to census matching, clerically reviewing over 200,000 candidate pairs to match all census records, is still too labour intensive. We are therefore unable to match the census to itself at a granular level, saying for each census record whether or not there is a duplicate record (at least not to the level of accuracy and timeliness that is required for estimation). Hence, we will have to estimate rather than count the overcount in 2021 Census.

We therefore propose to use the inverse sampling method of 2011 combined with the ACA described above. It is hoped that this will reduce the amount of clerical review required to around 27% of that required in 2011.

As in 2011, a probabilistic matching algorithm will be used to generate candidate duplicate pairs. This will be done by region and by groups who are most likely to include duplicates. In 2021, we will split London into two regions, Inner London and Outer London. There will therefore be 11 regions in total. In addition, candidate duplicates will have to match on names (allowing for some fuzziness), date of birth and sex. The method of finding duplicates is as follows:

- Split the population into ten regions.
- In each region, group people into one of 16 overcount-groups in hierarchical order.
- For every overcount-group and every region, take a random sample of 5,000.
- Candidate pairs are found from each sample through blocking on date of birth and sex, running probabilistic linkage and setting a threshold.
- Candidate pairs scoring above the threshold are sent through the ACA.
- If the number of candidate pairs automatically accepted plus confirmed matches sent to clerical review is 102 or above, stop the process for that overcount group and region.
- Otherwise[3], select another sample in that overcount group and region, adjusting the size according to the proportion of duplicates already found, and go through the process again until a total of 102 duplicates are found.

We propose a slight change to the groups deemed most likely to be overcounted. In 2021, we propose using the following 16 population groups, in priority order:

- Persons who have indicated they have a second residence on the census
- Female students aged 18 to 25
- Male students aged 18 to 25
- Persons who have indicated their address one year ago is not the same as their current address
- 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 completed the questionnaire on paper aged 16-29, 30-49 and 50+ (3 groups)
- Everyone else by broad age groups 16-29, 30-49, 50+ (3 groups)

3 If the number of duplicates found is less than 5 out of 5,000 for a particular overcount group and region then we stop as this overcount group has not been found to be more likely to be overcounted than the general population. If all the records in that overcount group for that region have already been classified then we stop.

These groups and their order have been determined by evidence from the longitudinal study [6] of overcount in particular populations, as well as the results from the 2011 Census to Census matching [7] and the results of our clerical review when developing the ACA.

The group, 'persons who have indicated their address one year ago is not the same as their current address', has been included because cases where address one year ago matched either the first or second address on the duplicate form, made up 7.8% of the duplicates found in the longitudinal study.

The group 'individuals who completed the questionnaire on paper' is included to primarily look for duplicates where one return is completed on paper and the duplicate is completed online. 5.6% of duplicates found in the 2011 longitudinal survey fell into the category of individuals who returned both an online and paper questionnaire at the same address. Although most of these (87.6%) were removed at the RMR stage of processing, it is worth keeping this category to catch further duplicates. Of the 455,889 candidates duplicates who were accepted automatically by the checking algorithm, 140,020 (31%) had completed using different modes.

As in 2011, we will use the hierarchical order of the groups and force them to be non-overlapping i.e., a person will be in the first overcount group that applies to them.[4] Thus, a child with a second address will be in the first group (second address). We accept that peculiar groupings may arise as a result e.g. communal establishment groups will not include any armed forces personnel. This may make it harder to produce estimates for the people who are in overlapping groups. However, this method is simplest to execute and understand and was implemented in 2011.

Conclusion

The census to census matching methodology for 2021 will make use of inverse sampling, together with the automated checking algorithm and thereby the amount of clerical resolution required for this matching exercise is expected to be reduced by over 70% without compromising on accuracy.

4 The only exception to this is students who say that they have an alternative address, but this alternative address is their out-of-term-time address. This group will be included in overcount groups 2 and 3 (students) rather than overcount group 1 (people with a second address). The reason for this is that the out-of-term-time address return for students is out of scope in the census and thus we would not expect to find a duplicate there. Such students are therefore no more likely than other students to be duplicated.

References

[1] Australian Bureau of Statistics, Census of Population and Housing: Details of Overcount and Undercount, Australia, 2016, available at: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2940.0~2016~Main%20Features~Contact%20Sector~21

[2] Methodology report on coverage matching for the 2021 Census, January 2019, ONS internal report, available on request or at: https://share.sp.ons.statistics.gov.uk/sites/MTH/Cen/2021/Data_Linkage_Methodology/Census_Censnus/Methodology%20report%20on%20coverage%20matching%20for%20the%202021%20Census.pdf

[3] Haldane, J.B.S (1945) On a method of estimating frequencies. Biometrika, Vol 33, No. 3, pp. 222-225

[4] 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

[5] Understanding the Birthday Paradox https://betterexplained.com/articles/understanding-the-birthday-paradox/

[6] Office for National Statistics (2014), 'Longitudinal Study 2011 Census Linkage Report', available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-userguide/quality-and-methods/quality/quality-assurance/index.html

[7] 2011 Census: Methods and Quality Report, *Overcount Estimation and Adjustment*, July 2012