Service leavers database linkage to 2011 Census

Methods used to link the Service Leavers Database, provided by the Ministry of Defence, to 2011 Census using deterministic and probabilistic methods.

Contact:
Hannah O'Dair

Last revised:
2 August 2023

Table of contents

1. Introduction to the linkage
2. Dataset quality and coverage
3. Methods
4. Quality information
5. Summary, recommendations, and limitations
6. Notes on deterministic matchkeys and Splink setup
7. Cite this methodology

Print this Methodology

Download as PDF

1. Introduction to the linkage

This report documents the linkage between 2011 Census and the Service Leavers Database (SLD). The SLD is a dataset provided by Ministry of Defence (MoD), containing the records of UK armed forces service leavers who left service between 1975 and 2022. However, for this linkage the SLD is restricted to those leaving service prior to census day 2011 (27 March 2011). This linkage and quality assurance is challenging because of limitations in the SLD variables available for linkage and differences in temporal and geographic coverage between the datasets.

Linkage between SLD and 2011 Census had been conducted before, as published in GOV.UK's Working age UK armed forces veterans residing in England and Wales bulletin. However, because of data retention policies, it had been deleted. The methods for this previous linkage were reviewed; however, it was decided the methods should not be replicated. Because of developments in the field of linkage, the methods previously applied would not meet current linkage best practice. As such, when re-designing a method to reproduce this linkage, more robust linkage and quality assurance methods were used.

Through a combination of deterministic and probabilistic methods, 40.5% of the deduplicated SLD is linked to 2011 Census. The linkage quality is deemed to be low, with precision between 60% and 96% (88% as the middle uncertainty tolerance estimate) and recall between 85% and 99% (96% as the middle uncertainty tolerance estimate). This means that even after clerical review, there is uncertainty whether a record pair represents a match or not, leading to a large range in the quality estimates.

Back to table of contents

2. Dataset quality and coverage

Data quality

The Service Leavers Database (SLD), as a dataset, has limited variables for linkage and has significant missingness in key variables. The variables useful for linkage were limited to forename(s), initials, surname, date of birth, sex and postcode. The level of missingness in forename and postcode (as reported in Table 1) provided a notable challenge for linkage. The missingness in these variables, together with the lack of other geography information, makes linkage on a population level very challenging.

By contrast, the census data had low missingness and was of good quality.

Table 1: Missingness of important linkage variables within 2011 Census and the SLD, UK, 1975 to 2011

	Percent missingness		
Variables	**SLD 1975 to 2022**	**SLD post deduplication (1975 to 2011)**	**2011 Census**
Surname	0.6%	0.7%	less than 0.1%
Forename(s)	15.6%	17.9%	lesss than 0.2%
Initials	1.5%	1.7%	N/A - derived from forename and surname
Date of birth	0%	0%	less than 0.4%
Gender	0%	0%	less than 0.5%
Postcode	62.1%	65.8%	0%

Source: Service Leavers Database from Ministry of Defence, 2011 Census from the Office for National Statistics

Notes

1. Initials on the SLD were provided by the Ministry Of Defence directly. For Census 2011, initials were derived from the given name variables.
2. Geography information is not held within the SLD and was sourced from the Compensation and Pensions System (Provided to ONS with the SLD by Ministry Of Defence). Not all veterans would have been in receipt of a UK armed forces pension or compensation, hence the low coverage.

Download this table Table 1: Missingness of important linkage variables within 2011 Census and the SLD, UK, 1975 to 2011

.xls .csv

Coverage

In addition to missingness, coverage differences between the SLD and 2011 Census make this linkage challenging. Firstly, the SLD data is historic (ranging from 1975 onwards), meaning there is a high chance that some persons will have died, or emigrated between 1975 and the 2011 Census (meaning they will not be present in the census), and there is higher chance of name changes (through marriage and divorce). Thus, historic data is harder to accurately link. Secondly, there is also a differing geographic coverage between the datasets, with census covering England and Wales and SLD covering UK armed forces service leavers, including those in Northern Ireland and Scotland. The temporal and geographic differences also mean that we do not know how many links are expected, between the datasets, which adds to the challenge of linkage and quality assurance.

Back to table of contents

3. Methods

Pre-processing

The Service Leavers Database (SLD) and 2011 Census underwent standardisation. Cleaning steps included case standardisation, removing punctuation and standardising date formatting. In addition:

- flags were created for records where names included titles
- a nickname variable was added to the datasets using the Office for National Statistics's (ONS's) nickname dictionary
- an indicator of unique biography was derived (where a full name and date of birth combination is unique within census)
- a military indicator (where a given census record contained an occupation or industry code that related to the military) was also derived

Deduplication of SLD

To deduplicate the data, three different rules were applied. The records with a service exit date after census day 2011 were removed prior to deduplication. For each stage of deduplication, the record with the most present linkage identifiers is retained, and then (where the number of identifiers agreed) the most recent record. Rules were applied in the following order:

- deduplicating on MODID, a variable provided to us by MOD which indicates the records belong to the same person for multiple periods of service
- deduplicating on forenames (including middle names), surname, full date of birth
- deduplicating on forename (no middle name), initials, surname, full date of birth, and postcode; this rule was added after finding that there were many apparent duplicates where middle name was missing

Following this deduplication, 1,603,782 person-level records remained (with service exit date between 1975 and 2011).

Deterministic linkage

The 2011 Census data is deterministically linked with the deduplicated SLD data using 27 matchkeys (details in section 6). Each matchkey consists of a set of rules or criteria that must be met to make a link. To account for expected errors in the data, the criteria are loosened on different linkage variables. Matchkeys were developed through trial and error, investigating the quality of links, and making iterative improvements. They are designed to account for transposed data, input errors and partial errors. Matchkeys are applied hierarchically, starting at the strictest matching criteria and becoming looser.

For cases where one census record linked to many SLD records or vice versa, the link with the lowest (first) matchkey was retained. This allowed the strongest link to be retained. Where conflicting links were made on the same matchkey, the links were both broken because of the inability to distinguish which is the correct link.

The result of deterministic linkage is a total of 611,066 matches with a match rate of 38% (of the deduplicated, pre-2011, SLD).

Probabilistic linkage

Probabilistic linkage was carried out on all records using Splink 2, a probabilistic linkage package, developed by the Ministry of Justice, which uses the Fellegi-Sunter method. Four local models were used to produce *m* and *u* values for each of the linkage variables (first name, surname, postcode, day of birth, month of birth, year of birth and sex). *m* values (agreement weights) are the probability that a variable agrees on two data sources given that they are a true match, so are a measure of data quality - how accurately the variable is recorded or freedom from error. *u* values (disagreement weights) are the probability that the variable agrees on both data sources given the pair are not a true match, so are a measure of distinguishing power or likelihood of matching by chance.

Probabilistic matching requires comparing each record on one dataset with all records on the other dataset to find a link. This results in a vast number of comparisons; the search space can be reduced by using blocking passes. Blocking passes mean only records which match on one or more specified variables are compared. This results in fewer comparisons being made, but those which are no longer made are ones unlikely to result in a true match. The blocking rules used in each local model are shown in Section 6. A global model was then constructed using the resulting *m* and *u* values and run to carry out the linkage.

Deduplication of Splink results was carried out, where the links with the highest match weight, for each census ID and SLD ID, were selected. Following deduplication, thresholds for acceptance of links were established by reviewing a small sample of records. Records with a rounded match weight of 24 (all of these had a match probability of greater than 0.98), were accepted; this decision was made in consultation with the client, balancing the tolerance for false positive against false negative errors. This resulted in 456,283 matches and a match rate of 28.5%.

Integration of deterministic and probabilistic results

The results of the probabilistic and deterministic linkage were joined together. Conflicting links were removed, as we were unable to determine which link was correct. This resulted in the removal of 4,886 conflicting links.

Following integration, there was a total of 649,186 linked records, with a link rate of 40.5%. Of the links, 413,277 (63.7%) were made both deterministically and probabilistically, 195,320 (30%) were made only deterministically and 40,589 (6.3%) were made only probabilistically.

Back to table of contents

4. Quality information

Clerical review

The standard approach to estimate error in the linked data is to perform clerical review (manual checking) on a sample of links and rejected record pairs. In linkage, there is a trade-off between two types of error - precision and recall. Precision (true positives divided by (true positives plus false positives)) is the proportion of the links made that are true matches, whereas recall (true positives divided by (true positives plus false negatives)) is the proportion of true matches that were found.

A sample of record pairs were run through the Data Linkage Hub's Clerical Review Online Widget (CROW) tool for review to clerically detect false positives (incorrect links, where a match has been made that should not have been made) and false negatives (missed links, where a match has not been made, that should have been) on a pair-wise basis.

As the service leavers data was limited, with a high proportion of the data not containing any geography information, the degree of uncertainty in clerical decisions was high. Therefore, the estimates of precision and recall are themselves limited in their accuracy. To capture this, a three-way review was conducted; each pair was reviewed by three different clerical matchers. The results were analysed to calculate best and worst-case precision and recall, as well as intermediate estimates. The clerical reviewers were all experienced and trained in clerical decision making and linkage. They were also briefed on the limitations of this data and given contextual information about both datasets.

For the clerical review, several extra variables (in addition to the name, date of birth, sex and postcode information used in linkage) were added to the data from census to aid decision making:

- address: the address string as on census
- unique biography: a flag to indicate whether, for the given census record, the full name and date of birth combination is unique within the census
- military indicator: a flag to indicate if census occupation and industry codes indicate the census record was a service member at the time of 2011 Census; matchers were instructed to use this as an indicator of a connection to the military, although were told not to give it too much weight
- name frequency: the standardised full name percentile frequency within census; it is between zero and one, with zero indicating extremely rare names and one indicating extremely common names on census

False positive review (precision)

A clerical review to estimate true positives (correct links, where a link has been made as it should have been) and false positives (incorrect links), is needed to estimate the precision of the linkage. For this review, pairs of linked records were grouped according to a combination of the deterministic matchkey they were matched on, and the probabilistic score the link obtained. The groupings can be seen in Table 2. Using these 15 strata, a total of 10,152 links were selected for clerical review. All links were reviewed three times and were grouped depending on the level of agreement between the clerical reviewers:

- 3/3 agreement of a match
- 2/3 agreement of a match
- 1/3 agreement of a match
- 0/3 agreement of a match

Table 2: False positive clerical review results, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

Matchkey	Probabilistic match weight	Strata	Total links in Strata	Number in sample	3/3 Match agreement	2/3 Match agreement	1/3 Match agreement	0/3 Match agreement

1 to 5	Greater than or equal to 38	1	58,733	750	749	1	0	0
1 to 5	30 to 37	2	92,018	750	739	9	2	0
1 to 5	24 to 29	3	260,44	750	704	46	0	0
1 to 5	No prob link	4	21,113	750	410	210	97	33
6 to 19	Greater than or equal to 38	5	4,480	750	742	6	2	0
6 to 19	30 to 37	6	56,013	750	693	45	7	5
6 to 19	24 to29	7	175,573	750	459	269	21	1
6 to 19	No prob link	8	170,518	750	127	428	149	46
20 to 27	Greater than or equal to 38	9	42	All-42	42	0	0	0
20 to 27	30 to 37	10	130	All-130	113	15	1	1
20 to 27	24 to 29	11	244	230	204	16	10	0
20 to 27	No prob link	12	3,689	750	70	181	263	236
No deterministic link	Greater than or equal to 38	13	1,223	750	700	37	11	2
No deterministic link	30 to 37	14	7,113	750	397	218	106	29
No deterministic link	24 to 29	15	32,253	750	154	142	209	245
Total			649,186	10,152	6,565	1,957	991	639

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this table Table 2: False positive clerical review results, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

.xls .csv

The classification of records where there was uncertainty depended on the precision bound being calculated. The methods used to calculate the worst case, intermediate and best case for the precision estimates are summarised in Table 3.

Table 3: Precision calculations split into best-case, worst-case, and intermediate estimates, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

	3/3 match agreement	2/3 match agreement	1/3 match agreement	0/3 match agreement	Precision
Worst case (lower bound) precision	True positive	False positive	False positive	False positive	60.1%
	True positive	True positive	False positive	False positive	88.1%

Intermediate estimate of precision				
Best case (upper bound) precision	True positive	True positive	True positive	False positive 96.3%

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this table Table 3: Precision calculations split into best-case, worst-case, and intermediate estimates, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

.xls .csv

The intermediate estimate of precision is 88.1%, where two-thirds of clerical reviewers agreed of either a match or non-match. However, our lower and upper precision estimate bounds are 60.1% and 96.3%, respectively. This range is large, indicating that our clerical review and consequently precision estimate, has a high degree of uncertainty.

False negative review (recall)

A review of true positives and false negatives (missed links) is needed to estimate the recall. For this review, record pairs from below the probabilistic threshold (non-links) were sampled by score region. A total of 7,000 record pairs were reviewed three times. Similarly to the false positive review, results were analysed based on the agreement between clerical reviewers. The results from both the false positive and false negative review were used to estimate recall and uncertainty ranges.

Table 4: False negative clerical review results, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

Probabilistic match weight	Group	Total links in strata	Sample	3/3 Match agreement	2/3 Match agreement	1/3 Match agreement	0/3 Match agreement
Less than or equal to 9	1	36,508	175	0	0	0	175
10 to 12	2	43,708	700	0	2	19	679
13 to 14	3	37,152	700	3	20	53	624
15 to 16	4	48,881	700	6	28	45	621
17	5	34,418	700	14	40	102	544
18	6	40,500	700	9	39	87	565
19	7	41,420	700	9	72	112	507
20	8	45,562	700	11	44	92	553
21	9	36,920	875	41	61	105	668
Greater than or equal to 22	10	44,639	1,050	33	56	172	789
Total		409,708	7,000	126	362	787	5,725

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this table Table 4: False negative clerical review results, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

.xls .csv

A total of nine recall estimates were calculated based on the three estimates of true positives from the false positive review and three estimates of false negatives from the false negative review, estimating recall for each combination of the best-case, middle-case and worst-case true positives and false negatives. The method for calculating the worst, intermediate and best-case recall estimate is shown in Table 5.

Table 5: Recall calculations split into best-case, worst-case, and intermediate estimates, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

	3/3 match agreement	2/3 match agreement	1/3 match agreement	0/3 match agreement	True positive estimate used	Recall
Worst case (lower bound) recall	False negative	False negative	False negative	True negative	Worst case true positive estimate	85.3%
Intermediate estimate of recall	False negative	False negative	True negative	True negative	Intermediate true positive estimate	95.7%
Best Case recall	False negative	True negative	True negative	True negative	Best case true positive estimate	99%

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this table Table 5: Recall calculations split into best-case, worst-case, and intermediate estimates, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

.xls .csv

The intermediate recall estimate was 95.7%, with the worst-case recall being 85.3% and the best case being 99%. This suggests that this linkage has identified between 85.3% and 99% of the links possible within our data.

Uncertainty summary statistics

Table 6: Summary statistics on the uncertainty of decisions in the different phases of clerical review, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

Clerical Phase	% Agree (across all 3 matchers)	% Do not agree (across all 3 matchers)
False Positive Analysis	71%	29%

False Negative Analysis	83.6%	16.4%
Across Both	76.1%	24.3%

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this table Table 6: Summary statistics on the uncertainty of decisions in the different phases of clerical review, SLD (UK, 1975 to 2011) and 2011 Census linked data asset

.xls .csv

Table 6 shows there is high uncertainty across both the false positive and false negative review. However, the level of disagreement (and thus uncertainty of decisions) was higher for the false positive review.

Figure 1 shows the uncertainty, broke down by review strata, for the false positive review (for each groups criteria, see Table 2). There was high level of uncertainty in groups 4, 8 and 12. This is notable, as these groups all matched on deterministic matchkeys but did not have a probabilistic match.

Figure 1: Percent of records where all three matchers agreed, by review strata in the false positive review of the SLD (UK, 1975 to 2011) to 2011 Census linked data asset

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this chart Figure 1: Percent of records where all three matchers agreed, by review strata in the false positive review of the SLD (UK, 1975 to 2011) to 2011 Census linked data asset

Image .csv .xls

Figure 2 shows the uncertainty, broke down by review strata, for the false negative review (for each groups criteria, see Table 4). The level of uncertainty was relatively consistent across groups, with groups that had the lowest match weight (lowest chance of being a match) having the highest consistency.

Figure 2: Percent of records where all three matchers agreed, by review strata in the false negative review of the SLD (UK, 1975 to 2011) to 2011 Census linked data asset

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this chart Figure 2: Percent of records where all three matchers agreed, by review strata in the false negative review of the SLD (UK, 1975 to 2011) to 2011 Census linked data asset

Image .csv .xls

Bias analysis

It is important to understand if there is linkage bias occurring within this data. Linkage bias is when the applied linkage method is better at capturing people with particular demographic characteristics, such that certain groups are under or overrepresented in the linked data. If unmitigated, linkage bias can lead to biased analytical conclusions.

In the absence of reference statistics, bias analysis is conducted by comparing the linked data with the source SLD data. However, for this linkage there is no accurate estimate of how many records are expected to link between the datasets. It is difficult to know whether differences between the linked and unlinked data reflects linkage failure (including because of data quality limitations) or differences in the coverage of the data.

Some records will not link because of legitimate reasons such as emigration and death, but it is hard to separate out these records from those which have not linked because of linkage error. Therefore, the interpretation of bias in this linkage is complicated.

To understand the potential biases in our linkage, proportional discrepancy was calculated for different variables. Proportional discrepancy is a measure of whether a particular demographic group is under or overrepresented in the linked data compared with the raw data. It is proportional to the overall match rate. It is on a scale of negative one to one; where negative one indicates severe under representation in the linked data, zero indicates proportional representation in the linked data, one indicates severe over representation in the data.

Year of exit and age

An analysis of bias within year of exit was performed, as shown in Figure 3. This shows an under-representation of people who left service between 1975 to 1980 and 1981 to 1986. All other groups are overrepresented in the linked data. This is likely to be because earlier data is of poorer quality and may be less likely to correspond with the 2011 Census. However, this trend could also be because of people having died or migrated prior to census day 2011. Whist these factors are indistinguishable, extreme caution should be taken when using year of exit as any observed patterns by year of exit, in analysis outcomes could be because of linkage bias; and may not reflect true trends or patterns.

Figure 3: Proportional discrepancy by year of exit, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this chart Figure 3: Proportional discrepancy by year of exit, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Image .csv .xls

Similar patterns were observed when investigating the bias by age group (Figure 4).

Figure 4: Proportional discrepancy by age group, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Notes:

1. Age used was age at census day 2011.

Download this chart Figure 4: Proportional discrepancy by age group, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Image .csv .xls

Rank

An analysis of bias by rank at exit date revealed a bias towards linking officers, as shown in Figure 5. The rank variable in the SLD denoted the rank at time of exit as either officers (OF1 to 10 including OFD) or other ranks (anyone who is not an officer, OR9 and below).

Other ranks were slightly underrepresented and missingness was highly underrepresented. Missingness may be underrepresented because of a higher likelihood of those records being of inadequate quality for linkage.

This indicates that caution should be used when considering rank within the data and that observed patterns by rank at exit date could be because of linkage bias rather than actual trends.

Figure 5: Proportional discrepancy by rank, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Download this chart Figure 5: Proportional discrepancy by rank, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Image .csv .xls

Rank by age group

As shown in Figure 6, an analysis of bias within other rank by age group revealed a bias towards linking other ranks aged 50 years and younger, and officers across all age groups aged younger than 83 years on census day 2011. Thus, there is an interaction between age and rank, such that people of other ranks, who were over 50 were notably under-represented. Officers in the same age group were overrepresented within the linked data (with the exception the over 83 group). In the under 50 age group, both ranks and officers were overrepresented in the linked data. Those with missingness in the rank variable were underrepresented across all age groups.

Figure 6: Proportional discrepancy by age group and rank, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Source: Service Leavers Database to 2011 Census linked data asset from the Office for National Statistics

Notes:

1. Missing values mean that there were no datapoints for the given group.
2. Age used was age at census day 2011.

Download this chart Figure 6: Proportional discrepancy by age group and rank, between SLD (UK, 1975 to 2011) and the SLD to 2011 Census linked data asset

Image .csv .xls

Sex

An analysis of bias by sex was conducted but is excluded from this report because of the low numbers of females in the data, leading to their removal from subsequent analysis.

Back to table of contents

5. Summary, recommendations, and limitations

In summary, 40.5% of service leavers database (SLD) records were linked to 2011 Census using deterministic and probabilistic methods. There were severe limitations to the data, and thus severe limitations to the linkage. Intermediate estimates of quality indicate 88.1% precision and 95.7% recall (with large uncertainty of these estimates). There was a bias by age and year of exit, with more recent records having a higher link rate. There was also some bias by rank.

Overall, caution is recommended when considering this linkage, because of the low link rate and precision. Analysts using this data need to be aware of the impact's linkage quality and bias, may have on their analytical findings. It is recommended that any resultant publications are transparent about the limitations of the linkage.

Back to table of contents

6. Notes on deterministic matchkeys and Splink setup

Deterministic matchkeys

Notes for matchkey variables:

- full name refers to the combination of all given first, middle and last names
- unique biography is a flag on the census data that indicates that a given full name and date of birth combination is unique within census

- everworked is a flag within the census data that indicates that a person has at some point in their life, been in employment
- the forename variable was split on space; creating forename1, forename2 and forename3 plus
- Jaro-Winkler is a method for assessing the similarity between string variables

List of matchkeys

- MK1: concordant full name, date of birth, sex, and postcode; they must have a unique biography and have ever worked within census
- MK2: concordant forename, surname, date of birth, sex and postcode
- MK3: concordant initials, surname, date of birth, sex and postcode; they must have ever worked within census
- MK4: concordant full name, date of birth and sex; they must have a unique biography and have ever worked within census
- MK5: concordant forename1, forename2, surname, date of birth, and sex; they must have a unique biography within census
- MK6: Jaro-Winkler > 0.9 on forename1; concordant surname, date of birth, sex and postcode
- MK7: Jaro-Winkler > 0.75 on forename1; concordant surname, date of birth sex and postcode
- MK8: Jaro-Winkler > 0.6 on forename1; concordant surname, date of birth, sex and postcode
- MK9: Jaro-Winkler > 0.9 on surname.; concordant forename1, date of birth, sex and postcode
- MK10: Jaro-Winkler > 0.75 on surname; concordant forename1, date of birth, sex and postcode
- MK11: Jaro-Winkler > 0.6 on surname; concordant forename1, date of birth, sex and postcode
- MK12: Jaro-Winkler > 0.7 on postcode; concordant forename1, surname, date of birth and sex
- MK13: Jaro-Winkler > 0.5 on postcode; concordant forename1, surname, date of birth and sex
- MK14: Jaro-Winkler > 0.9 on forename1; concordant initials, surname, date of birth and sex. Must have ever worked with census
- MK15: Jaro-Winkler > 0.9 on surname; concordant forename1, initials, date of birth and sex; Must have a unique biography within census
- MK16: concordant forename1, surname, date of birth, and sex; must have a unique biography and have ever worked within census

- MK17: concordant forename1, surname and date of birth. Must have a unique biography within census
- MK18: SLD nickname concordant with census forename1; concordant surname, date of birth and sex and they must have a unique biography, and have ever worked within Census
- MK19: census nickname concordant with SLD forename1; concordant surname, date of birth and sex, and the must have a unique biography and have ever worked within census
- MK20: transposed day and month of birth; concordant forename1, surname, birth year and sex and the must have a unique biography and have ever worked within Census.
- MK21: missing day of birth; concordant forename1, forename2, surname, month of birth year of birth, and sex and they must have a unique biography and have ever worked within census
- MK22: missing month of birth; concordant forename1, forename2, surname, day of birth, birth year and sex and they must have a unique biography and have ever worked within census
- MK23: missing year of birth; concordant forename1, forename2, surname, day of birth, birth month and sex and the must have a unique biography and have ever worked within census
- MK24: transposed forename1 and forename2; concordant surname, date of birth and sex.
- MK25: Jaro-Winkler > 0.7 on forename 1;concordant surname, date of birth and sex, and they must have a unique biography and have Mil_Ind flag (indicating the census record has a sic or soc code indicative of military occupations)
- MK26: Jaro-Winkler > 0.7 on surname; concordant forename1, forename2, date of birth and sex and they must have a unique biography and have Mil_Ind flag (indicating the census record has a sic or soc code indicative of military occupations)
- MK27: Jaro-Winkler > 0.7 on surname; concordant forename1, forename2, date of birth and sex and they must have a unique biography within census

Splink setup

The blocking rules used in each local model are set out as follows:

- local model one: equal initial of first name, surname and year of birth
- local model two: equal postcode
- local model three: equal date of birth and first name
- local model four: equal surname, day of birth and month of birth

These values taken from each of the local models to be used in the global model are as follows:

- local model one: day of birth, month of birth, and sex
- local model two: year of birth
- local model three: surname and postcode

- local model four: first name

Back to table of contents

7. Cite this methodology

Office for National Statistics (ONS), released 2 August 2023, ONS website, methodology, Service leavers database linkage to 2011 Census

Back to table of contents

Contact details for this Methodology

Hannah O'Dair
linkage.hub@ons.gov.uk

Footer links

Help

- Accessibility
- Cookies
- Privacy
- Terms and conditions

About ONS

- What we do
- Careers
- Contact us
- News
- Freedom of Information

Connect with us

- X
- Instagram
- Facebook
- LinkedIn
- Consultations
- Discussion forums
- Email alerts