Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Scotland's Census 2022

Item level Imputation Using Administrative Data – Date of Birth

August 2020

National Records of Scotland

Contents

1. Plain English Abstract

Date of birth is a crucial piece of information collected in the census. This is because it is used to calculate a person's age, which helps us understand the age profile of the population. This paper considers presents the method for dealing with Census returns with dates of birth that are missing, or obviously incorrect through inconsistencies with other census variables. This is done by linking such census records to an administrative data source. Where a record matches an administrative data record, the age on the admin record is used to help find a similar census record that has age recorded. That census age is then used as the age for the problematic census record.

2. Abstract

Date of Birth is a crucial piece of information on the census form, it is used to derive age at census day, which is one of the most important variables in the census. This is required for the main statistical outputs for census statistics.

This paper examines the proposed methodology to deal to cases where the date of birth data is missing from the census return. This could be in error, information being unreadable on scanned forms. It also looks at cases where the date of birth on the census return is obviously incorrect. For example, this could be because the date of birth indicates that the respondent is a child, but also indicates that the respondent is married and is employed.

Census records are linked to an administrative data source. The linking considers administrative data records in the same postcode as the census record in question, and compares them on first, middle and last names, sex, and date of birth if available.

Where a record matches an administrative data record, the age on the administrative data record is used, along with other variables on the record, to identify a suitable donor record from among the census returns. The age recorded on the donor record may then be used in place of the age for the record with missing or inconsistent date of birth.

Note: On 17 July 2020 Scottish Government announced the decision to move Scotland's Census to 2022 following the impact of the COVID-19 pandemic. The information included in this report reflects the methodology intended, at the time of publication, to be used in the 2022 Census. It is not expected that there will be any major differences between the methodology presented here and that used in the 2022 Census. However, some detail may change or be completed before or during census processing. Any major changes to the intended methodology will be described in an update here.

3. Introduction and background

Sometimes a census return will not include responses for all questions. Some questions may not be relevant for all respondents (such as marital status for children). However, in other cases question responses may be blank where they should have been completed. This could be deliberate non-response (for example due to perceived privacy concerns), or it could be that a question was skipped by mistake, or because the respondent thought that it was not relevant to them. With online returns if a respondent completes part of the form then this data will be collected as an unsubmitted return. In such cases, responses to questions toward the end of the form may be missing. With paper returns a response could be illegible.

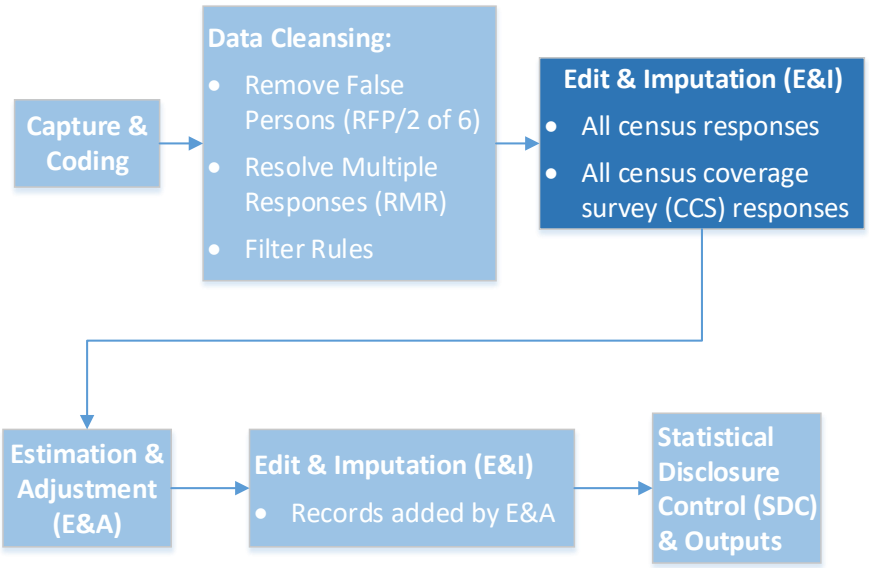

Figure 1 Where imputation fits into Statistical Data Processing.

In general, such missing data is dealt with using hot-deck imputation. This is where, for each record (A) with missing data, another record (B) is found from among the census responses that has similar data to record A on questions that have been completed. Then the value that record B has for the question that is missing for A is used to complete the question in record A. Record B is then referred to as the donor record. Figure 1 shows where imputation is carried out in the data processing

sequence. Imputation is described in more detail in the Overview of Edit and Imputation methodology paper[1].

In addition, census responses can occasionally be incorrect. This may be due to errors respondents make when inputting data, or it could be due to errors made when paper returns are scanned. When such errors are detected then hot-deck imputation is used to modify the data to make it plausible.

Date of birth (DoB) is used to derive age at census day, one of the most important census variables. It is used to produce the high-level breakdown of the population, and is a common filter for other census outputs (such as those for working-age people). Having accurate age information can also help with detecting and correcting other problems with the census data (such as parental relationships being the wrong way round).

To address this issue The Office for National Statistics have explored assisting the hot-deck imputation process by adding age from a linked administrative dataset (Leather, Sharp and Rogers, 2018). For the 2022 Census in Scotland a similar process is proposed. To do this, census records will be linked to an administrative dataset. Links will be made using information on name, postcode and sex. If present, date of birth will be used, although this will not need to be exactly the same. The links will be used at the imputation stage. However, the linking itself takes place during the remove false persons step, as administrative data linking is also used there. This means that the linking happens before census records are resolved in RMR.

If a match is found in the administrative data for a census record with missing or incorrect date of birth, then the age from the matching administrative data record can be used to inform the age for the record. The administrative data age will be used to assist in the hot-deck imputation process, meaning that it is more likely to find a

[1] See NRS (2020a).

donor record with similar age. This paper describes the proposed method for linking to administrative data to find matches, and using them in imputing the age.

The 2022 census will be primarily collected online, generally respondents who do not wish to respond online will request a paper questionnaire. In the online questionnaire, respondents are shown the age corresponding to the date of birth they have entered. This gives respondents an opportunity to notice and correct any errors, improving overall quality. Also, respondents will not be able to submit an online questionnaire without entering the dates of birth, improving completeness.

However, some online questionnaires may remain unsubmitted without date of birth information, but will still be collected. Also the aim is for 80 per cent of returns to be online, which would still leave around a million responses on paper. Therefore, while it is expected that age information in 2022 should be improved from 2011, it remains likely that a number of records will have problems with age information and so need resolved.

Section 4 describes what was done in 2011. The proposed 2022 method for linking the census to the administrative data is described in Section 5. The testing of this linking method is given in Section 5.4. Section 5.4 then describes how the information from the administrative data will be used to address issues with age information.

4. 2011 Method

In 2011 age was included in the demographics module and imputed using hot-deck imputation. The other variables in this module were sex, marital status, full-time student, term-time location, relationships and economic activity. Hot-deck imputation would then find a donor record that was similar on these variables. The age of that donor record would then be used for the record requiring imputation. Administrative data was not used. The problem with this method is that these variables are relatively weak predictors of age compared with linked administrative data.

5. Proposed 2022 Method

For 2022, it is proposed that administrative data are linked to the census. The administrative data age then assists the imputation. All the census records are linked to the administrative data source, using components of the linking method developed for the Census to CCS linking methodology[2]. The linking method is described in Section 5.1. The way duplicate cases are dealt with is discussed in sections 5.2 and 5.3. The proposed method was developed using 2011 census data linking to an administrative data source. It was then tested on the 2019 census rehearsal data, the results of which are given in sections 6.2 and 6.3. Once the administrative data age has been added to the census record it is used in imputation to assign an age. How this is done is described in Section 5.4.

5.1 Linking Census to Administrative Data

Every census record is compared with every administrative data record in the same postcode (blocking[3] on postcode). For each pair of records a score is calculated for the evidence for the pair being a match and the evidence against them being a match, for particular variables. These variables are first, middle and last names, and sex. In the census questionnaire name is collected using two fields: forename(s) and surname. The information in surname is used for last name. The first part of forename is used for first name, and the remainder of forename is used for middle name. The comparisons on name can account for nicknames, phonetically similar names, and typographical or scanning errors. Further detail on this is in Annex 1.

[2] See NRS (2020b).

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

When date of birth is present on the census record, this will also be used for linking. Exact agreement is not required on date of birth, but some level of similarity is. For example, the records might have the same day and month of birth but a different year.

These links are then categorised based on the for and against scores for each linking variable, with each category having an associated distance score. Links with a distance score of 0 or 1 were deemed strong enough to be accepted automatically, those with distance scores from 2 to 6 (or where the distance score is 7 and it is classed as a possible parent–child pair) were passed for clerical review, and those with a distance score of 8 or 9, along with other categories with distance scores of 7, were not considered. These thresholds were developed in order to closely mimic the judgements of clerical review. In addition, a sample of automatically accepted links are taken for clerical review for quality assurance purposes. For further information on the process and scoring see annexes 1 and 2.

As the process does not insist on identical dates of birth, there will be cases where multiple records link together. This will be a particular issue when there are parent–child pairs who have the same name and live at the same location. Some further steps are therefore taken to address such problems. When date of birth is missing, where this will be a particular problem, is covered in Section 5.2, while the equivalent for when dates of birth are not missing is covered in Section 5.3.

5.2 Handling Duplicate Cases: Missing Date of Birth

A common problem when linking without date of birth is that parents might link to their children. This would happen if there is a parent and child with the same name at the same address. Suppose that the child has missing date of birth, and that the parent (but not the child) appears on the administrative dataset, as in Table 1. As the name and postcode agree a link would be formed between the child on the census and the parent on the administrative dataset (record C2 and A1). In such

cases the link between the census child and the administrative parent (C2–A1) should be broken so that the parent's age is not used for the child's census record.

Table 1 Some (fictitious) records illustrating the need for administrative data records to be linked back to other census records.

Dataset	Record	Name	Age	Postcode	Note
Census	C1	Maya Patel	42	AB1 1AA	Parent
Census	C2	Maya Patel	Missing	AB1 1AA	Child
Administrative data	A1	Maya Patel	42	AB1 1AA	Parent

Alternatively the two census records may represent the same person. This might happen if a household begin completing a census form and enter the names of the people in the household but do not complete all the individual forms (and so the date of birth will be missing). If the respondents went to complete the form later, but had forgotten their password, then they would have to complete a new form. Both these forms are collected, so there ends up being two census records for the same person, one with missing date of birth, as in Table 2. As before the link between the record with missing date of birth would get broken and the administrative age would not be used. Although that record would continue through processing without an age, this should not be a problem. At the Resolve Multiple Returns step such cases would resolve the two census records into one record, and the date of birth would be used from the record where it was completed.

Table 2 Some (fictitious) records all representing the same individual illustrating the need for administrative data records to be linked back to other census records.

Dataset	Record	Name	Age	Postcode
Census	C3	Olivia Wilson	35	G1 1AA
Census	C4	Olivia Wilson	Missing	G1 1AA
Administrative data	A2	Olivia Wilson	35	G1 1AA

To check each link, the records from the administrative dataset are then linked to the full census dataset (not just records with missing date of birth). This linking is done in the same way as described above. In general, if census with missing date of birth links to multiple administrative data records, or if the administrative data record it links to also links to other census records then the link will not be used. However there are two exceptions. The first involves cases when two census records with missing date of birth both link to the same two administrative data records. Such cases are likely to be parent–child pairs. If the census records come from the same questionnaire and there is information from the relationship matrix then it would be possible to identify that they are a parent–child pair, and to match the records to the parent and child. If the ages on the two administrative data records differ then the older administrative age can be assigned to the census record representing the parent and the younger administrative age to the one representing the child.

Another exception is when a census record with missing date of birth links to two administrative data records, but one of them links to another census record that has date of birth included. So if a parent and child had the same name and had records both on the census and the administrative data, and only one of these records had missing date of birth, then this is what is likely to happen. In such cases the administrative data record that does not link to the census record that includes date of birth can be used for the census record that does not have date of birth.

The rules about dealing with multiple links are summarized in Table 3. The steps used to apply these rules are given below. Table 3 indicates the number of the step that addresses each of the particular rules.

1. Load and prepare data
2. Link Census records with missing date of birth to administrative data records
3. Trim the set of links down to the strongest link for each census record, along with any other links for that record that are within a distance score of 2 of that strongest link.
4. Link the linked administrative data records to census records that do have date of birth recorded
5. Remove from the main set of links any links involving administrative data records that linked to census records that have date of birth

6. Identify cases where two census records with missing date of birth link to the same administrative data records. Retain cases where the two census records are a parent–child pair and the administrative data records reflect this
7. Remove from the main set of links any links with records involved in the cases identified in step 6
8. Remove from the main set of links any links where the administrative data record links to multiple census records (with missing date of birth)
9. Remove from the main set of links any links where the census record links to multiple administrative data records and the administrative data records have different ages
10. Bring together the main set of links with those identified in step 6

Table 3 Rules for handling multiple links, with each rule's justification.

Situation	Action	Justification
1 census missing DoB record and another census record both link to the same admin record	Do not use the link. Step 5	The admin record likely matches the other record. The missing DoB record is either a non-match or will be resolved out at RMR anyway.
2 census missing DoB records both link to the same admin record	Do not use the link. Step 8	The age might be assigned to the wrong record[4].
1 census missing DoB record links to 2 admin records with different ages	Do not use either link. Step 9	There would now be certainty that either particular age was correct.
2 census missing DoB records link to 2 admin records with different ages	Do not use the links, unless one census record is the parent of the other, and the ages reflect this. Step 6	Unless we had relationship information then we would be <50% sure that either age was correct.
1 census missing DoB record links to 2 admin records, but one of the admin records links to a census record that includes DoB	Use the link to the admin record that does not link to the other census record. Step 5	It is likely that the other admin record matches the other census record, leaving a 1–1 link between the census and admin records.

[4] There would be the possibility that the date of birth linking could be rerun after RMR. This may then detect cases where two census records with missing date of birth had been resolved into one.

5.3 Handling Duplicate Cases: Date of Birth not Missing

When date of birth is not missing, such problems are less likely to occur as linking also makes use of the date of birth information. Children may still link to their parents records, but it will generally be clear which record they should have linked to. Therefore there are relevant two steps for when dates of birth are not missing:

1. Remove cases where the census record links to multiple administrative data records with the same distance score. For each census record, only keep the link with the best distance score.
2. Remove links where the administrative data record links to another census record with the same or better distance score.

Once links have been removed using the above two rules each census record, and each administrative data record, will not appear more than once.

5.4 Imputation method used

Once the records have been identified and an administrative data age attached, the records are passed to the data processing team where the imputation technique is applied. This involves six steps, as described below.

Step 1. An extra column 'admin age' is added to the census dataset, for administrative data age.

Step 2. The administrative data age is copied into the 'admin age' column, only where there is an administrative data link to census data and one of the following conditions is met:

(i) Census age is missing but administrative data age is available

(ii) Census age is different to administrative data age

(For example, records 2 and 4 in Table 4)

Step 3. For empty spaces in the 'admin age' column, if there is a census age for that record, it is copied into the 'admin age' column
(Records 1 and 3 in Table 4)

Table 4 Example records from **before** imputation.

Record ID	Census Age	Admin Age	Other variables…
1	23	23	…
2	missing	25	…
3	56	56	…
4	0	51	…

Step 4. Records are flagged as 'failed records' which need to be imputed, if they contain inconsistencies, invalid values, or missing values.
(Record 2 in Table 4 are missing census age and so need to be imputed. Record 4 in this example has an inconsistency with some other variable not shown, such as marital status)

Step 5. For each 'failed record', the automatic software searches for similar records, using the given characteristics[5]. This includes the 'census age' and 'admin age' columns. Where 'census age' must be imputed, the 'admin age' column helps to find records of a similar age to the admin age. The imputed value comes from the 'census age' column of the donor record. Sometimes the imputed value will be the same as the 'admin age' in the failed record, and sometimes it will be different, but usually quite close.
(In Table 4, the donor for record 2 is record 1, and the donor for record 4 is record 3.)

[5] The other characteristics in the demographics module are: marital status, country of birth (in/outside UK), economic activity, full-time student, relationships, term-time location, age, sex, enumeration location, response mode (paper/online) and hard-to-count code.

Table 5 Example records from **after** imputation. Cells shaded orange indicate where the census age has been imputed.

Record ID	Census Age	Admin Age	Other variables…
1	23	23	…
2	23	25	…
3	56	56	…
4	56	51	…

Step 6:

An imputation flag is added and the Admin Age column is removed from the statistical dataset before statistical disclosure control processing is implemented. The admin age is attached to a temporary copy of the dataset in order to export to CANCEIS[6], and with the current CANCEIS settings only variables which are imputed will be written out after imputation. The variable will always be in those input files, and it will always be in one of the audit trail files which is produced, but the actual census dataset which is in the 'core dataset' store never has to see it.

In summary, any census record where the age is missing or inconsistent with other variables (such as being the age of a child but married) will be imputed, and age may be changed as part of this imputation (see Table 6). This is the case whether or not a link was found to an administrative data record. Any census record that has a valid age and passes the edit is a potential donor. Again, these may or may not have been linked to administrative data. For each record needing imputed a record from the donor pool is selected that is similar on the imputation variables, including admin age. The census age of the record needing imputed is then set to the census age of the donor record.

[6] A piece of off the shelf software from Statistics Canada which is used to undertake imputation. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2009/wp.15.e.pdf

Table 6 Breakdown of census records and how they are used.

Census records may have an age available, or it may be missing. The census record will fail the edit if age is missing or inconsistent with other information. For each census record, a matching administrative data record may or may not be found. If there is a match then the ages could be the same or different, if there is not a match then the census age is used for the admin age.

Census age		Administrative data age		Role in
Available?	Consistent?	Linked to census?	Same as census	imputation
No	N/A	Link found	N/A	Record imputed
No	N/A	No link found	N/A	Record imputed
Yes	No	No link found	Yes by definition	Record imputed
Yes	No	Link found	Yes	Record imputed
Yes	No	Link found	No	Record imputed
Yes	Yes	No link found	Yes by definition	Potential donor
Yes	Yes	Link found	Yes	Potential donor
Yes	Yes	Link found	No	Potential donor

6. Results

This section presents results from testing done using 2011 census data and 2019 census rehearsal data. Section 6.1 presents some initial feasibility results using 2011 census data. This explored the relative quality of the census and administrative data ages, by linking with the census coverage survey (CCS). It also explored how well hot-deck imputation could recover the administrative data age when using it as an imputation variable.

The remainder of this section presents results of testing the methods for linking and resolving issues with duplicate records for missing date of birth (Section 6.2) and not missing date of birth (Section 6.3), using the 2019 census rehearsal data. The National Health Service Central Register (NHSCR) was used as the administrative data. This is a dataset of persons born in Scotland or registered with an NHS GP in Scotland.

6.1 Previous Analyses Using Census 2011 Data

Previous research linked the census 2011 data to dataset formed by combining the NHSCR with a health activity dataset. To assess the relative quality of the age data of the census and the administrative data cases where the dates of birth differed on the census and administrative data were identified. These cases were then linked to the census coverage survey. It was discovered that the CCS age agreed with the administrative data age more often than it agreed with the census age. This suggested that overall the administrative data is more reliable than the census data for date of birth.

The test on the 2011 data also explored how close hot-deck imputation with and without administrative data ages would get to using administrative data ages directly. Figure 2 shows that in many cases the age assigned when using administrative-data assisted hot-deck imputation is the same as if the administrative data age had been used directly.

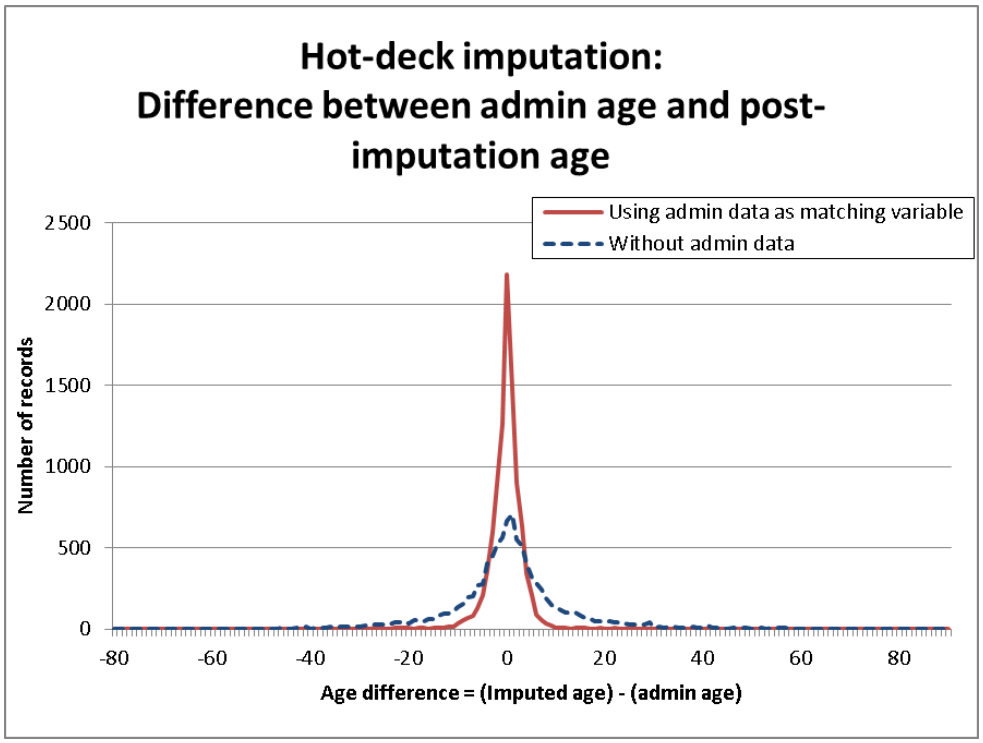

Figure 2 Distribution of the difference between the imputed age and the administrative data age when using two methods of hot-deck imputation when testing on 2011 census data.

Table 7 Number and proportion of cases in the 2011 test where the imputed age was within a given range of the administrative data age, both when using the administrative data age, and without using it.

Age difference	Number of cases		Percentage of cases	
	With admin data	Without admin data	With admin data	Without admin data
Exact year	2,187	669	21.8	6.7
Within 1 year	4,928	1,944	49.1	19.4
Within 5 years	9,081	5,390	90.4	53.7
Within 20 years	9,833	8,871	97.9	88.3
Total	10,042	10,042	100.0	100.0

In addition, Table 7 shows proportions of the imputed ages that are within a certain distance of the administrative data age. When the administrative data age is used in

the hot-deck imputation the ages used are close to the administrative data age itself, than when using hot-deck imputation without administrative data age. For example, in around half of the cases the age is within a year of the administrative data age, compared with around a fifth when not using administrative data age. Hot-deck imputation with the administrative data age therefore gives results that are closer to using the administrative data age directly, than to the results when imputing without administrative data.

6.2 Rehearsal: Date of Birth Missing

In total there were 7,594 records in the rehearsal dataset that had a missing date of birth, once the person based dataset had been cleaned. 6,943 (91.4 per cent) of the cases were from online returns and of these they were all 'unsubmitted returns'[7]. Many of these cases will be when respondents began completing the questionnaire and entered the names of the household members, but did not reach the individual questionnaires where dates of birth are entered. The remaining 651 (8.6 per cent) were from paper forms.

6,675 links were found between these records and the administrative data. These links involved 5,716 distinct census rehearsal records. There were 740 administrative data records in these links that also linked to census records that did have date of birth (Step 4). Removing links involving such administrative data records reduces the number of links to 5,205 (Step 5). From among these links, 30 census records were involved in parent–child pairs where the two census records linked to the same two administrative data records with sufficiently different ages (Step 6).

[7] Online returns which were not submitted by the respondent, but some data had still been entered.

Table 8 Rehearsal records with missing date of birth linking to NHSCR by category of the strongest link, and whether the rehearsal record came from an online or paper return. This shows the numbers after all the filtering for duplicate cases has been completed.

Category of strongest link	Online returns	Paper returns	Total
0 Exact	0	20	20
1 Same (A)	4,200	126	4,326
2 Same (B)	22	2	24
4 Likely same (A)	73	30	103
4B Name same, miss DoB	51	9	60
5 Likely same (B)	22	16	38
6 Likely same (C)	68	19	87
Total	4,436	222	4,658

Following the removal for further cases in steps 7 and 8, 4,346 of the records with missing date of birth linked to the administrative dataset strongly enough that the link could be automatically accepted (that is, with a link that was categorised as either '0 Exact' or '1 Same (A)'), see Table 8. 207 of these were reviewed as a matter of course to quality assure the linkage programme and all were passed by the reviewer. A further 312 records linked but needed clerical review.

Table 9 shows the age breakdown of the linked NHSCR records. This shows that there is a higher proportion of such cases with age less than 30 than in the general population and a lower proportion for older age groups. This might be expected, given that many of such cases were from online returns, which may be a more popular mode of responding among younger people. It also shows that if these ages were imputed from other census responses then, depending on what other information was available to guide imputation, that could lead to a systematic error in tending to select records for older persons.

Table 9 Rehearsal records with missing date of birth linking to NHSCR by the 5-year age band on NHSCR.

Age	Number of cases	Percentage of cases	Percentage of population
0–4	345	7.4	5.5
5–9	436	9.4	5.1
10–14	450	9.7	5.5
15–19	381	8.2	6.2
20–24	395	8.5	6.9
24–29	321	6.9	6.5
30–34	267	5.7	6.1
35–39	256	5.5	6.4
40–44	234	5.0	7.5
45–49	245	5.3	7.8
50–54	290	6.2	7.1
55–59	269	5.8	6.2
60–64	217	4.7	6.4
65–69	137	2.9	4.9
70–74	165	3.5	4.2
75–79	125	2.7	3.4
80–84	68	1.5	2.4
85–89	39	0.8	1.4
90+	18	0.4	0.6
Total	4,658	100.0	100.0

Thus if these results were scaled up to a full census, this would equate to roughly 500,000 cases where a date of birth from administrative data would be supplied which would aid the imputation process, including around 30,000 cases to review. It should be noted, however, that in live census this number could be notably lower. Recall that in situations where a respondent forgets their password and completes a new form the link would be broken (as the administrative record would also link to another census record). In the voluntary census rehearsal, respondents who forget

their password might be less likely to phone up for a new form, than they would in the compulsory census. Thus in census more of the individuals with missing date of birth records might also have records with completed date of birth. In these cases the link to the administrative dataset would get broken, and the two census records would be resolved together at the Resolve Multiple Responses step. Until the compulsory, predominantly online census is carried out, it is impossible to quantify how much of an effect this will have.

6.3 Rehearsal: Date of Birth not Missing

The methodology for linking and dealing with duplicate cases was also tested on the 2019 Census rehearsal data for date of birth not missing. There were 37,961 census rehearsal records where date of birth was not missing. 31,146 of these linked to the NHSCR (see Table 10). Once the cases involving links to multiple records are removed this reduces to 30,327.

Table 10 Number of census rehearsal records that do not have date of birth missing that link to the NHSCR, broken down by the source of the response. This is shown both before and after the records have been filtered due to duplicates.

Response source	Number of records	
	Before filtering	After filtering
Online unsubmitted	4,540	4,153
Online submitted	23,913	23,574
Paper	2,693	2,600
Total	31,146	30,327

In total there were 328 cases where the age differed between rehearsal and admin data (see Table 11). Of these, 258 were paper and 70 were from online returns. It might be expected that respondents would be more likely to make mistakes with this on paper forms, or for them to be due to scanning errors. With the online returns

people have to enter into the web based form and then it takes the date of birth and presents them with their age, as a built in quality assurance check. This offers the respondent a chance to change this themselves if it did not look right. If the age displayed is over 114 years old, or born after census day then the respondent will be displayed with a validation message asking them to confirm or alter the details.

Table 11 Census rehearsal records after filtering by source of the response and the difference in the age between that recorded on the census rehearsal record and the NHSCR. Cases where the census age is not available are where the date of birth is incomplete or invalid.

	Same age	Census age different from admin age		Census age not available		Total
		same 5-yr band	different 5-yr band	age known within 1 year	age unknown	
Online unsubmitted	4,142	9	2	0	0	4,153
Online submitted	23,515	42	17	0	0	23,574
Paper	2,342	27	65	11	155	2,600
Total	29,999	78	84	11	155	30,327

All the cases where the administrative data age is the same as the census age (29,999 in the test) would be passed for automatic acceptance. 201 of the remaining 328 cases would be automatically accepted, and 127 would be passed to review (see Table 12).

Table 12 Cases with different ages between census rehearsal and NHSCR by whether it would be passed to clerical review and the difference in age. Cases where the census age is not available are where the date of birth is incomplete or invalid. If the year of birth is available then the age would be known within one year of the true age.

Review status	Census age available. Difference between census and admin ages		Census age not available		Total
	different age, same 5-yr band	different 5-yr band	age known within 1 year	age unknown	
Auto	66	50	7	78	201
Review	12	34	4	77	127
Total	78	84	11	155	328

These records can be compared with the records where age is inconsistent with other variables. All of the 155 cases where the census age was not available would fail the edit. The 11 cases are where an exact age could not be derived, but the age was one of two adjacent ages. This might be if month of birth is missing or invalid. In these cases one of the two possible ages would be chosen at random.

Table 13 Census rehearsal cases where the census age is available by whether census age is inconsistent with other information on the census, and whether the census age is the same as the administrative data age.

Census age inconsistent with other information	Comparison between census and admin ages		Total
	same	different	
Yes	66	3	69
No	29,933	159	30,092
Total	29,999	162	30,161

A further three of the remaining 162 cases are where there is a different age on the administrative data record and the census rehearsal record had an inconsistency

involving age (see Table 13). All three of these were from paper returns. These are the cases where the record would be imputed, and the administrative data age would assist in the imputation (see Section 5.4 for more detail on this). There are 66 cases with an inconsistency on age, but the administrative data age is the same. Providing the administrative data age in such cases will provide evidence for CANCEIS to preserve the age and change the other information in order to resolve the inconsistency.

Cases where there is not an inconsistency on age, no changes will be made by imputation. NRS are exploring flagging up the remaining paper returns with different administrative data ages in order for them to be manually checked against the scanned images. This would allow scanning errors to be detected, but saves the effort of having to look through the scanned images for all paper returns.

If this is scaled up to the full 2022 census, for the total population, this would indicate there would be approximately 20,000 cases where the admin data age could be potentially used in the imputation method, around 400 of which would actually get imputed, making use of the administrative data age. We are considering only reviewing cases where the administrative data age would be used in imputation (if the link was not flagged for automatic acceptance). Cases being flagged for checking against the scanned image do not need to be reviewed, as the checking against the image is itself the review. However, whether the link was flagged for automatic acceptance would be considered when prioritising which cases to check, along with the magnitude of the difference in ages.

7. Strengths and Limitations

7.1 Strengths

Date of birth, and the age on census day derived from it, are very important census variables. Age is often used as an analysis variable for the census (for example when presenting the age distribution of the population). It is also often used as a filter (for example when investigating the attributes of particular age groups). Linking to administrative data adds an additional layer of quality assurance to the process that was not present in the 2011 census.

In the 2011 census, cases where there was no date of birth on the census form had to be imputed using other information provided by the respondent, such as household relationships, to find a suitable donor record. Including the age from a matching administrative data record provides further information to direct and constrain the imputation.

In the cases where administrative data is matched to give additional intelligence to select an appropriate donor record, there is a clear improvement on the 2011 method. Previous analysis using the 2011 CCS discovered that the CCS age agreed with the NHSCR age more often than it agreed with the census age, suggesting that overall the NHSCR data is more reliable that the census data for date of birth. Because of this, and the successful test using NHSCR with the 2019 census rehearsal data, it is planned that the NHSCR will be used as the administrative data source for 2022.

7.2 Limitations

Limitations are around the amount of resource needed to deliver the required clerical review. However, it may be possible to substantially reduce the amount of clerical review required, by only reviewing cases where the census record fails the edit. The other limitation is that the process has not yet been implemented in an end to end

test, where it fits in the overall sequencing and it has not been integrated into the data processing system. However, the intention is that these would be worked on and any issues ironed out before the census live run. These though are not issues with the methodology itself, just the application and delivery of it.

8. Conclusion

Overall this method provides additional quality assurance, not just to the variables it is providing but to the overall process further downstream in data cleansing of Scotland 2022 Census. It offers a clear improvement on the 2011 census as imputation method will be nearer to the individual's true age. While resource outlay could be considered a factor against the adoption of this methodology, it would be a modest and prudent investment to ensure a product with much improved data quality for such a widely used variable.

9. References

Leather, F., Sharp, K. and Rogers, S. (2018) 'Towards an integrated census - administrative data approach to item-level imputation for the 2021 UK Census' Neuchâtel

National Records of Scotland (2020a) 'Overview of Edit and Imputation' [Online] available at: https://www.scotlandscensus.gov.uk/documents/Scotland's%20Census%202022%20-%20EMAPs%20-%20PMP012%20Overview%20of%20Edit%20and%20Imputation%20for%202022%20(pdf)(2).pdf

National Records of Scotland (2020b) 'Census to Census Coverage Survey (CCS) linking' [Online] available at: https://www.scotlandscensus.gov.uk/documents/PMP010_-_Census_to_CCS_linking_-_EMAP.pdf

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

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

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

10. Annex 1: Scoring of Name Comparisons

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

Missing Names

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

For first names there is also a check for the name being 'BABY' on both records. In this case the for and against scores are both set to 0 as the guidance (in 2011) indicated that unnamed infants should be recorded as 'BABY'. This scoring was developed for the Resolve Multiple Responses step, where two BABYs could easily be twins. In the Census Coverage Survey linking the likelihood may be different so this may need revised.

Nicknames

Another check for first names is nicknames. Thus if we had 'Alexander' on one record and 'Sandy' on the other then it is quite plausible that these are the same person, even though the first name strings are quite different. To perform this check we make use of the nickname linking variable. That variable is set to a particular value for a range of names that have the same nickname. Thus if first was either 'Alexander' or 'Sandy' (or 'Alex', 'Xander', and others) then the nickname variable is set to 'Alexander'. (The name groupings were built up manually, assisted by exploring links between datasets where last name, date of birth and postcode

agreed, but first name did not.) Thus if the first names differ between records but the nicknames agree then the against score is set to 0 and the for score is set to 20. Some of these are specific to a particular sex. Thus if the first name is 'Alex' then the nickname will be set to 'Alexander' if sex is male and 'Alexandra' if sex if female. There is also a second nickname variable that groups together more tenuous name groupings such as 'John' and 'Ian', which results in a for score of 10.

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

Character comparison for names

If none of these situations hold then the name components in the two records are compared at the character level using a method inspired by the Damerau–Levenshtein edit distance[8]. The characters in the name from one record are linked to those in the name from the other record. This is done by first comparing the characters at the same location in the strings. If these do not agree then this moves to adjacent letters, and then letters at a distance of two, and so on. Once this has completed there is a tidying up stage to ensure that adjacent letters are linked to letters at the same distance if possible.

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

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

Combinations like 'U' and 'V' can be easily confused, as can 'O' and 'D'. In total 50 such combinations are noted.
The scores from all the individual differences are then combined to give an overall score. That score is then converted to scores for and against the records being a match.

Swapped first and last names

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

Titles

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

Comparison to middle name

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

Compare name parts

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

Comparing first letters of name or Double Metaphone code

The next check is to count the number of letters that agree at the start of the name from the two records. If so then the for score is set to be that given in Table 14. This covers a range from one letter agreeing to five (or more) letters agreeing. If only one letter agrees then this is treated differently, so that this method is used only if one record only has the initial (e.g. if one record had 'Peter' and the other had 'P', but not if the other was 'Paul'). These scores are only used if they result in a higher for score than would otherwise be. Another exception is when 3 or fewer letters agree and the names are distinct but common. For example if we had Mary and Margaret then the first three letters agree, but as the names are common then this is not used to score the similarity.

Table 14 The for scores assigned when the first part of the name agrees either on the name itself, or the Double Metaphone coding of it. If only one letter agrees then this method is only used if one of the records only has one letter.

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

Similarly the first characters of the Double Metaphone are compared. The Double Metaphone is a phonetic code[9], so this allows for detection of cases where a name has been written differently, but sounds the same. This is another situation that may be particularly common for verbally reported data such as the Census Coverage Survey. As a character in the Double Metaphone code can relate to more than one letter in the original string, agreement on Double Metaphone can indicate stronger agreement than agreement with the same number of letters on the original string. Therefore these scores are slightly larger than the equivalents for the agreeing letters on the original name.

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

Full name

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

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

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

better than the for scores for first and last name then the first and last for scores are amended using the full name for score.

11. Annex 2: Scoring of Sex and Date of Birth

Sex

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

Date of Birth

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

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

Table 15 Sets of digits that may be confused in scanning, and so are given a smaller difference penalty.

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

If the two dates of birth are complete then the individual digits are compared. That is, the first digit of the day of birth from one record is compared with the first digit of the day of birth from the other record, then the second digit and so on. If the two digits are both in one of the sets given in Table 15 then we count this as a difference of 1. All other differences are counted as a difference of 2. (The particular sets of digits are chosen to be those that are often confused in scanning, so are more likely to be the same than for other pairs of digits.) These differences are then totalled across the whole date of birth.

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

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

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

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

12. Annex 3: Categorisation of Links

Once the for and against scores have been calculated for each component for each link, the links are placed into one of the categories shown in Table 16 below.

Table 16 List of categories used to class the links along with a brief description of the condition used to place them and the strength associated with the category. The categories are presented in order of the priority in which they are assigned. That is, links are only assigned to a given category if they do not meet the conditions for any preceding categories.

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

Distance score	Name	Description of Condition
7	Don't know last partial agree	As Don't know first partial agree but with condition on last
7	Likely different	Total for >50, total against <20
7	Probably different	Weak evidence against from first, last or DoB, total for > total against
8	Different – sub	Weak evidence against from up to two of first, last and DoB
9	Different other	Evidence against from first, last and DoB
7	Remaining	Any records not assigned to any of the above categories

13. Annex 4: Glossary

Term	Definition
Link	Two records that have been connected
Match	Two records that represent the same individual
Non-match	Two records that represent different individuals
NHSCR	The National Health Service Central Register. A dataset of persons born in Scotland or registered with an NHS GP in Scotland.
CCS	Census Coverage Survey. A survey carried out independently of the census to estimate the coverage of the census.
RMR	Resolve Multiple Returns. A data processing steps where multiple census records relating to the same individual at the same location are identified and resolved into a single census record.
Edit	The detection of missing, invalid or inconsistent responses.
Imputation	The correction of missing, invalid or inconsistent responses.
Donor	A record that is used to help impute a failed record. Response values are copied from the donor to the failed record in order to replace missing or inconsistent responses.
Hot-deck imputation	Donor records come from the same dataset as the failed record.

14. Annex 5: Information Governance

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

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

Data Protection Impact Assessment for use of NHSCR dataset

Quality Assurance report for use of NHSCR dataset for 2019