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

Scotland's Census 2022

Census–CCS Address Linking

February 2021

National Records of Scotland

Contents

1. Plain English Abstract

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

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

This paper looks at how we decide whether addresses in CCS and census records are the same or not. This can help inform decisions on whether person records represent the same person. It will also be used to identify which households respond to both the census and CCS, which is used to estimate how many households are missed by the census.

2. Abstract

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

This process relies on an accurate count of the people and households appearing on both sources. In order to achieve this it is useful to determine whether a CCS record has the same address as a census record.

The first step to do this is to standardise the addresses to correct common address shortenings, remove irrelevant characters and to use consistent address naming conventions where possible.

Each CCS address is then compared to census addresses using a variety of matchkeys with a link between addresses being recorded if there is exact agreement between the matchkeys. The first groups of matchkeys find the most obvious links where a significant portion of the standardised address (such as property, street and postcode, or property information, street and town) are exactly the same. If a link between a CCS address and census address is found with any of these groups of matchkeys then the CCS address is removed from the pool of addresses that need to be linked.

If a link cannot be found using that group of matchkeys then we look for links for the remaining addresses using matchkeys that consist of more limited portions of the address, for example only selecting the numbers in the address.

Once all of the comparisons have been made, the set of links are collated into one dataset. For most CCS addresses this dataset will contain a link to a single census address, however in other instances the CCS address will link to more than one census address. This dataset of address link information then feeds into the wider estimation process with the address links being clerically reviewed where necessary as part of the Census–CCS person/household linkage.

3. Introduction and Background

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

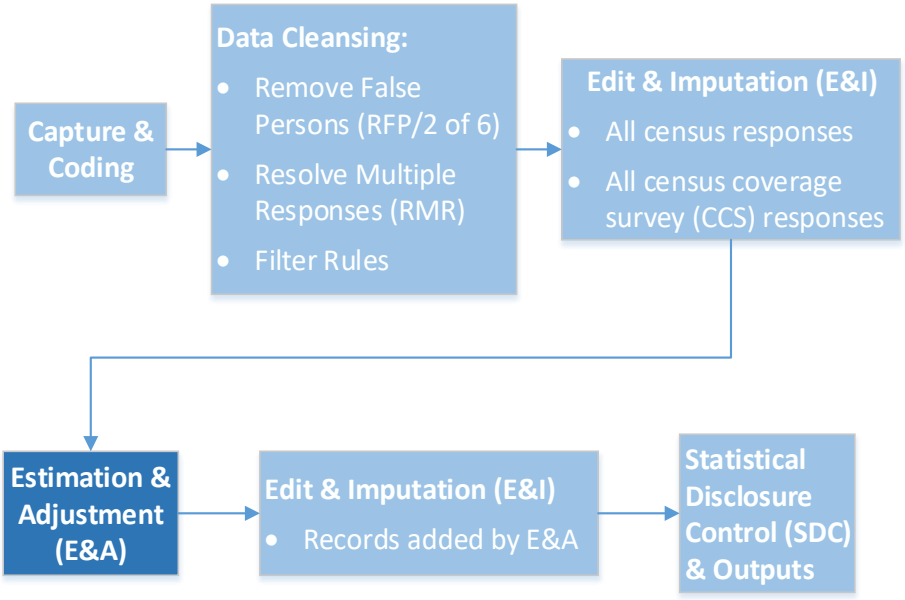

Figure 1 Where estimation fits into Statistical Data Processing[1].

An accurate population estimate relies heavily on accurate linking between the CCS and census records. The CCS is a sample of around 1.5 per cent of postcodes. Thus each match that is not linked will increase[2] the population estimate by around 80. Conversely each link that should not have been made will reduce the population

[1] See NRS website on Statistical Methodology (2020) for more information on each of the stages within Data Processing for the 2022 Census.

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

estimate by around 80. Therefore the impact of a single error at this stage has a much larger impact on the final census outputs than individual errors at other processing stages, such as Remove False Persons and Resolve Multiple Returns[3] (where a single error would typically affect the estimate by around 1).

As well as estimating the number of people in Scotland, the census also estimates the number of households in Scotland. The census definition of a household is:

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

This definition is taken to mean that a household is both the people **and** the address. Therefore, a household link will be formed between census and CCS if a return is received for the same address (that is, an address that is matched) in both the census and CCS, with at least one corresponding person link.

Identifying whether addresses match is not always straightforward as the address frames for the CCS and census are independent of each other so addresses will not always be recorded in a consistent manner. Some of these inconsistencies are minor and straightforward to resolve, for example if one address is 'Flat 1, 12 High Street' and the other is '1, 12 High Street'. However some are more complicated, for example if parts of the address have been omitted or there are differences in spelling. Another problem is the use of different naming conventions to refer to the same flat ('1F1, 12 John Street' being the same as 'Flat 3, 12 John Street' or 'Flat C, 12 John Street' or '12/3 John Street'). This can be particularly problematic as there is not a universal mapping between the various formats, and it is impossible to predict which format will be used as it can even vary between buildings on the same street.

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

This paper looks at how links between CCS and census addresses are determined. Detail on how the wider estimation process[4] and Census–CCS Person Linking[5] are covered in a separate papers.

[4] Estimation and Adjustment Methodology https://www.scotlandscensus.gov.uk/documents/Scotland%E2%80%99s%20Census%202022%20-%20PMP001%20-%20Estimation%20and%20Adjustment%20Methodology%20(pdf).pdf.
[5] Information will be published on our website: External Methodology Assurance Panels (EMAPs) | Scotland's Census

4. 2011 Method

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

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

The linking variables used were:

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

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

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

It was therefore decided to develop a new methodology, that could be audited better. This included a more comprehensive address matching method where the full address information is used rather than just an address token.

5. Proposed 2022 Method

Address linking methodologies have been pursued across the UK with the Office for National Statistics (ONS) using data science for address parsing and creation of an Address Index (AI). The process of splitting up an address, or tokenisation, can be complicated because people use different formatting and structures for their address. The process is further complicated by nature of the human element in deciding how to label different words of an address in AddressBase, the reference dataset used by ONS[6].

The address linking method in this paper is a by-product from a project to produce household estimates purely from administrative data, where addresses from an administrative dataset are linked to the Scottish Address Directory, the reference dataset used to create the Census Address Register (CAR). The primary administrative dataset used to develop the code for this method was the Health Activity 2016 dataset. While the administrative data used to develop the code will not be used in any way during the CCS to Census address linkage, the code developed to match addresses is still relevant as the address data in the Health Activity dataset is more similar to what we expect in the 2022 Census than other sources. The other sources that we considered using were address data from the 2011 Census and the 2019 Census rehearsal. However the majority of addresses in 2011 were provided on paper forms which poses a different set of challenges, while the addresses in the 2019 largely came directly from the Census Address Register so would be identical.

Our decision to use the bespoke response to address matching for CCS to Census address matching was also influenced by local needs, including handling of many rural dwellings with unusual addresses that may contain Gaelic spellings, and a high number of tenement properties where flats are numbered in quite varied formats.

[6] ONS working paper series no 17 - Using data science for the address matching service

5.1 Method Summary

The methodology for linking addresses has the following features:

- It is written in SAS, which is readily available in NRS, reducing the need for licences and training in other applications, and can be reviewed easily by other statisticians.
- The process uses matchkeys that are split into 5 groups based on the rationale behind them. The matchkeys are concatenations of the different variables that make up each address, such as building number and street name. For some matchkeys these variables are modified to identify further links.
- If the matchkeys do not find a link then a comparison of property names for all addresses within the same postcode is made in an attempt to find addresses that are similar enough that we believe they are the same address.
- CCS addresses are compared to both responding census addresses and non-responding census addresses.

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

1. Standardise addresses
2. Generate matchkeys
3. Link using matchkeys in Groups A and B — exact match on at least property name and postcode, or property, street and postcode
4. For addresses not linked in the previous step, use matchkeys in Groups C and D — exact match on property, street and town
5. For addresses not linked in the previous steps, link using matchkeys in Group E — match within postcode using property numbers in the address only
6. For addresses not linked in the previous steps, link using fuzzy matching on house names within the same postcode
7. Collate all links that have been found so this information can be used in the wider estimation process.

5.2 Method Detail

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

Step 1: Standardise addresses

Census addresses are taken from the Census Address Register (CAR), which is a cut of the Scottish Address Directory (SAD) held by NRS. Those census addresses use a standard format with the following fields:

- Organisation
- Property
- Building Number
- Street
- Locality
- Town
- Postcode

In contrast, the CCS addresses are collected by enumerators (field force). Those staff register every dwelling in the postcodes allotted to them. Field force employees register the address on a mobile device using the following fields:

- Establishment name (for communal establishments only)
- House name or number (addressline1)
- Street (addressline2)
- Addressline3
- Town (addresstown)
- Postcode

This means the addresses may not be an exact match to those identified in the CAR. There is also a question in the CCS where respondents are asked if they lived at a different address on census day. Where an address is provided as an answer to this question, that address is used rather than the address in the variables above. This address is in a different format from the above, recorded in just two variables, one for the address and one for the postcode. The address linking process for these

addresses is therefore slightly different and described in Annex 2. However this should only apply to a small number of addresses.

Once all of the address data is available, each element of the CCS and census addresses, with the exception of postcode, is standardised, with the standardised address being stored as a new variable. The aim of this is to create modified variables where common variations of the same information are standardised to the same value. This means that common minor differences in addresses are removed making it is easier to find a link between addresses. Two versions of standardised address variables are produced, one with minimal changes and another with an increased level of standardisation.

For the version where minimal changes have been made the standardisation consists of:

- changing all characters to upper case.
- removing special characters with the exceptions of - and / as they are characters that are an important part of many flat naming conventions.
- spaces before or after a - or / are removed.
- standardising the word 'AND' to '&'.
- SAINT changed to ST.
- expanding common abbreviations for street names, for example, 'AVE' to 'AVENUE', 'RD' to 'ROAD' and 'ST' to 'STREET'.
  To minimise erroneous changes such as 'ST ANDREWS' changing to 'STREET ANDREWS' an underscore is introduced in the middle of town names containing the word 'ST' (e.g. ST_ANDREWS).

Matches identified using this minimal standardisation or 'cleaning' of the address variables, are made with a high degree of confidence. This first wave of matching was suggested as desirable by our Geography team who valued having the resultant address variables being as close as possible to how they were originally recorded.

For the second, more comprehensive, standardisation of the address variables, the following changes are made in addition to those above:

- The word 'THE' is removed as this is not a distinguishing feature of an address and is sometimes omitted (for example, 'Old School' instead of 'The Old School')
- Any words beginning with 'Mac' are changed to 'Mc'
- Any instances of a number followed by a dash are changed to a number followed by a slash (for example, 13-2 STEWART CRESCENT changes to 13/2 STEWART CRESCENT)
- Common flat naming conventions where a '-' is included are changed to slashes (for example, 'G-F' is changed to 'G/F' or '13-2' changed to '13/2')
- The words 'FIRST', 'SECOND', 'THIRD' are changed to '1ST', '2ND', '3RD'
- Common abbreviations for 'FLOOR' and 'FLAT' are changed to the full word
- Common descriptions of flats based on the floor, and whether it is on the left or right side are standardised (for example, 'GROUND FLOOR RIGHT' and 'GROUND RIGHT' are both changed to 'G/R')
- Common descriptions of flats based on the floor it is on, without reference to a side are standardised (for example, 'GROUND FLOOR FLAT' and 'GROUND FLAT' are changed to 'G/F')
- After the above is done then words beginning 'G/' are changed to '0/' and those ending '/L' and '/R' are changed to '/1' and '/2' respectively[7]
- Common variations of COTTAGE, FARMHOUSE, HOUSE, and LODGE are changed to the full word
- The words NORTH, NOR and NTH are changed to N. Similar changes are made for SOUTH, EAST and WEST
- All dashes are removed. As the dashes in flat names have been changed to slashes, these can now be removed as they can cause confusion when used in a house name
- The word 'FLAT' is removed to allow links to be made where the word 'FLAT' has been omitted.

[7] This is based on a clerical review of such addresses and residential flat numbering conventions published by several Scottish councils. For example Street Naming and Numbering Conventions (renfrewshire.gov.uk).

In some cases this will mean the cleaned address does not necessarily look as it is ever would in real life. For example 'Ground Floor Flat, 14 MacLaren Dr' is possibly never written as 'G/F 14 MCLAREN DRIVE'. However this is not an issue when the changes are made consistently across both datasets, and do not cause two different addresses to become indistinguishable.

Step 2: Generate linking variables

After the variables have been standardised the matchkeys are generated. These matchkeys are created by concatenating combinations of the address elements and/or selecting parts of the address within each element that are of interest (for example any numbers within the address).

The overarching rationale for groups A to F set out below, is to start with the maximum address data available and then to consider dropping components of the address that clerical review highlighted as being potentially problematic. Once a unique match was identified, subsequent explorations were only deployed on unmatched addresses and so on until the maximum number of matches could be secured. For instance postcodes were noted as being wrong or incomplete for a number of addresses, so matchkeys were generated without the postcode variable to ascertain whether the remainder of the address produced a match. Similarly some properties had variant spellings, including Gaelic or local interpretations, rendering it unlikely to secure a match with all text included in the matchkeys. This led to investigating the success of matching only numeric components of an address combined with postcode, in a similar way to websites where house number and postcode is often sufficient to uniquely identify an address. In an attempt to pick up remaining unmatched addresses we considered 'close fits' through a scoring mechanism used in character comparison for names, as documented in Census-CCS Person Linking (NRS, 2020), with the fuzzy match strategy as our final group. The sequential approach led to groups of matchkeys being explored, first with minimal cleaning and then the more comprehensive standardisation of address variables. The matchkeys are split into groups based on the rationale for the creation of each matchkey.

Steps 3 and 4 describe the rationale behind each group of matchkeys and illustrate the types of addresses they manage to link. The full list of the matchkeys is provided in Annex 1.

Step 3: Link using matchkeys in Group A and B

The matchkeys in Group A and B identify the most obvious links where a significant part of the standardised addresses match exactly. There are two run-throughs for each of these groups, the first using the minimally standardised address variables and the second with the more fully standardised versions. The order for this is:

1. Group A with minimally standardised address variables
2. Group B with minimally standardised address variables
3. Group A with fully standardised address variables
4. Group B with fully standardised address variables

The general criteria for a link to be found with these groups of matchkeys is provided below and a full description of the matchkeys is provided in Annex 1.

Group A: This group of matchkeys identify links between addresses when the following conditions hold:

- The CCS address and census address have the same postcode.

And either:

- The property and street appears in both addresses or;
- The house name appears in both addresses, provided the house name is not just a number.

While these are the minimum criteria for a link to be found, the matchkeys include the different combinations of the other information, such as town, as well. The reason for this is that in some cases information from the street variable in the CCS may be found in locality/town in the census dataset or vice versa. This is particularly likely in rural areas where a street name is not relevant, but something must be entered in this field for CCS addresses.

Group B: This group is a modification to the matchkeys used in Group A that helps to identify additional links where the only difference is due to additional spaces in either the CCS or census address.

To do this an underscore is inserted between any numbers in the address that are separated by a space, and then all spaces are removed. Otherwise the minimum requirement for Group A still applies.

This allows additional links to be found where either the CCS or census address has an extra space (or a space removed). For example if the CCS had an address recorded as HILLSIDE FARM but the census has HILL SIDE FARM.

Step 4: Link using matchkeys in Group C and D

For all CCS addresses where a link to a census address was not found in the previous step we move onto the next groups of matchkeys. These groups identify links between addresses where the postcodes do not match, but there is clear evidence that they are the same address. Incorrect postcodes should be less of an issue with CCS and census addresses than the administrative data used to develop the address linking method as there will be more address validation. Therefore it is not expected that many links will be found with these groups, however these matchkeys have been retained to provide extra assurance that the links found are correct.

As with matchkey groups A and B, there are two run-throughs for each of these groups, the first using the minimally standardised address variables and the second with the more fully standardised versions. The order for this is:

1. Group C with minimally standardised address variables
2. Group D with minimally standardised address variables
3. Group C with fully standardised address variables
4. Group D with fully standardised address variables

The general criteria for a link to be found with these groups of matchkeys is provided below and a full description of the matchkeys is provided in Annex 1.

Group C:
For this group of matchkeys the following conditions are required for a link:

- The house number/name, street and town from the CCS address appears in the census address as well.
- Postcodes do not have to be equal

Group D: This group modifies the matchkeys in Group C in exactly the same way as Group B modifies Group A. An underscore is introduced between numbers which are separated by a space and then all spaces are removed. As with Group B, this allows some additional links to be identified when there are only minor differences in the addresses due to additional spaces.

Step 5: Link using matchkeys in Group E

For any addresses where a link was not found in the previous steps a slightly different approach is used. For many addresses the key features that make the address unique are the flat/building number and the postcode, and the rest of the address can effectively be ignored. This group of matchkeys take advantage of this by extracting any component in the address that contains a number, or a flat identifier (these were standardised to end with a '/F' in step 1) from the address. One of the advantages of this approach is that any spelling errors will be ignored so will not affect whether a link is made.

Three examples of this are shown below, with the full address and the matchkey that is produced. The last example demonstrates how typos or alternative spellings in an address would not prevent a link being found as the same matchkey is produced as the correct version of the address.

Address	**Matchkey**
7/3 Broad Street, Edinburgh, EH9 9ZZ	7/3 EH9 9ZZ
Flat 16, 25 High Street, Scotland, KW15 2XY	16 25 KW15 2XY
Flat 16, 25 Hihg Street, Alba, KW15 2XY	16 25 KW15 2XY

There is an additional step where the address is searched for any occurrences of 'Flat X' where X is any letter. If this is the case then the letter is appended to an appropriate component of the matchkey where possible.

Address	Matchkey
Flat A, 32 Ronaldsay Avenue, Perth, PH1 1SY	32A PH1 1SY
32 Ronaldsay Avenue, Flat A, Perth, PH1 1SY	32A PH1 1SY
Flat D, Pierowall Place, Inverness, IV2 2TS	D IV2 2TS

This group also contains matchkeys which modify addresses containing a slash by separating it into its components. This allows links to be found where the CCS has flat numbering written in a slash format and the census is in more long form, or vice-versa.

For example, if the CCS address is '7/3 Broad Street, Edinburgh, EH9 9ZZ', two matchkeys are created '7 3 EH9 9ZZ' and '3 7 EH9 9ZZ'. This allows a link to be found if the address is recorded as 'Flat 7, 3 Broad Street' or 'Flat 3, 7 Broad Street' in the census. Both possibilities are included as the convention for whether it is the building number or flat number before the / appears to be inconsistent across Scotland. If both Flat 3, 7 and Flat 7, 3 exist then both links are recorded and we have two potential links for the CCS address.

Step 6: Link using fuzzy matching on house names within the same postcode

The matchkeys in the previous step took advantage of numbers in addresses, however not all addresses contain numbers. This step attempts to find a link for any addresses where a link has not been found already and does not contain any numbers.

Unlike the previous steps, this does not involve matchkeys. Instead the house name of each CCS address still to be linked is compared to the property information for all census addresses where the postcode matches the CCS postcode.

This comparison is made using a string comparison algorithm which was developed as part of the name linking process. The algorithm produces two scores that

measure the similarity between two strings. The first of these scores is based on the longest string of consecutive characters that the two strings have in common. The second is based on the number of substitutions, deletions, insertions, transpositions and jumps required to convert one string into the other. All links found where these scores are below a certain threshold[8] are then recorded.

Step 7: Collate all links

Once the search for links is complete, a dataset containing all of these links is created. Each row consists of a unique CCS–Census address pair. If a CCS address has been linked to multiple census addresses then it will appear on multiple rows. Information about which matchkey group made the link between the addresses and whether the link found is a one-to-one link is added as a new variable.

The dataset of links is then ready to be incorporated into the wider CCS–Census linking of persons and households.

For person linkage, the address links can assist by providing additional evidence to support whether CCS and census records are for the same person. Fuller details of how this is likely to impact on person linking will be developed jointly by Admin Data and Data Processing teams.

The address links will have a more vital role in judging whether a household is present in the CCS and census as the address must match as well as there being at least one person in common. Therefore some clerical review may be required to ensure the accuracy of household links. The exact process for determining when clerical review will be required is yet to be finalised, however circumstances where it may be necessary include situations such as:

- Incorrect links. Any incorrect links will be identified once the address links are combined with person link information. If there is a conflict between the person and address links then it will suggest that further investigation is required.

[8] The thresholds are determined following a clerical review so that the vast majority of recorded links are correct

- A CCS address linking to multiple census addresses. If a person is found on the CCS and census but the address link is only one of many that it linked to then clerical review may be used to provide assurance that the addresses are the same.
- CCS addresses not linking to any census records. If there is no address link then clerical review will be required before a household link can be confirmed.

6. Results Using Addresses From Administrative Data

6.1 Testing plan

The results from testing the method come from the development of the method for use in a project to produce administrative data based population and household estimates from administrative data where address linkage was also required.

It was not possible to use the 2019 Census Rehearsal for testing as it did not include a CCS. Using address data from the 2011 Census and CCS for testing would be another possibility, however CCS addresses were collected on paper and scanned, while in 2022 the majority of addresses will be collected electronically. This means that challenges in linking the data are significantly different as scanning error is more prevalent and the structure of addresses is not the same. For this reason the testing results presented are based on linking a health activity dataset (referred to as the development dataset) and Scottish Address Directory (SAD). We believe the address data in that development dataset is a good reflection of what we might expect in 2022, and the address methodology required for that project, could be used for Census to CSS address matching. .

The SAD is used to create the Census Address Register which will be the source of the majority of census addresses, so the differences between this and the final census addresses will be minimal. However there are some differences between the addresses in the development dataset and those that will be collected in the CCS. The main differences are listed below and should be taken into account when considering the results of this testing and thinking about them in the context of CCS to census linkage.

1. While addresses in the development dataset are separated into different fields, what is contained in each field is not defined. In other words we cannot say that the second field of the address will contain street information as we can with the CCS.

Impact/Mitigation

An alternative method of identifying the street name was used where necessary, for example in CCS_MK_A1[9] which includes property information, street and postcode. Instead of just selecting the appropriate variables, to replicate this with the development dataset the full address string was searched for common street signifiers such as street, road, avenue and lane. The matchkey then consisted of the address string up to and including the last appearance of a street signifier plus the postcode. In the majority of addresses, this will results in the same matchkey being produced. However in some instances, particularly in rural areas, an address may not have a street signifier in the address so the matchkey ends up being the full address string.

2. There is less validation of the addresses in the development dataset, particularly for postcodes. As a result there are more addresses with an incorrect or invalid postcode in the development dataset than there will be in the CCS that is structured and selected on the basis of known postcodes.

   Impact/Mitigation

   There is little that can be done to mitigate this, so instead the impact of this on the results has to be acknowledged. Firstly, having more addresses with incorrect postcodes will increase the number links found using matchkeys from Group C and D. Additionally there will be an increased number of incorrect links made using matchkeys in Group E, as it will find links to any property in the incorrect postcode with the same number as the address being linked. The higher level of validation on CCS postcodes would mean that we would expect more links to be made using Group A matchkeys instead.

3. There is not a separate Establishment Name variable for communal establishments.

   Impact/Mitigation

   The fact that there is no separate Establishment Name has to be largely ignored in this testing and is only relevant to communal establishments. In the development dataset the establishment name is either included in the

[9] See Annex 1

address, usually at the beginning of the address, or not at all. In any case, the establishment name is only used to create variants of the matchkeys by its inclusion/exclusion and therefore should not have a large impact.

The linkage method was performed on 3,639,583 unique address strings from the development dataset compared with 2,804,101 address strings in the SAD[10]. After running the linkage method a random sample of the links were clerically reviewed to check whether incorrect matches are being identified or not. The addresses that were not linked to an address in the SAD were also clerically reviewed to identify the reasons for a link not being made.

6.2 Results

A link between each address from the development dataset and a single SAD address was found for 86% of addresses, as shown in Table 1. However, this percentage should be higher when linking CCS addresses to census addresses due to some of the characteristics of the addresses in the development dataset that could not be linked. These characteristics are discussed later in this section.

Table 1: Breakdown of addresses from the development dataset by the number of SAD addresses they are linked to

Number of addresses linked to in SAD	Number	Percentage
One	3,131,232	86.0%
Two or more	34,675	1.0%
None	473,676	13.0%
Total	3,639,583	100.0%

Table 2 shows the distribution of which matchkey groups identified the links for the addresses that were linked to one address in the SAD. As expected, the majority of the links were identified with matchkeys from Group A.

[10] There are fewer records in SAD compared to the development dataset because multiple individuals may reside at the same address and some addresses are duplicated within this if there is some difference in how the address has been captured. For example: 5 Main Road, KW10 2TN and 5 Main Rd, Highlands, KW10 2TN are different address strings for the same address.

Table 2: Breakdown of which matchkey groups identified the link for those addresses that linked to one address in the SAD

Matchkey Group	Number of links	Percentage of all links found
A with minimally standardised addresses - at least property and postcode	2,503,219	79.9%
B with minimally standardised addresses - spaces removed from Group A	38,573	1.2%
A with fully standardised addresses	233,346	7.5%
B with fully standardised addresses	4,251	0.1%
C with minimally standardised addresses - property, street and town the same but not postcode	28,642	0.9%
D with minimally standardised addresses - spaces removed from Group C	2,254	0.1%
C with fully standardised addresses	52,700	1.7%
D with fully standardised addresses	1,447	0.0%
E - only considers numbers and postcode must be the same	233,898	7.5%
Fuzzy matching	32,902	1.1%
Total	3,131,232	100.0%

A random sample of 1,000 of the linked addresses were reviewed to see how many incorrect links were made. Of the 1,000 links, 993 were obviously correct, 2 were possibly correct but there were plausible alternatives and 5 were definitely incorrect.

For the two possibly correct links:

- Both were farms/cottages where the link is likely to be correct, but there are plausible alternatives. For example, Hillhead Farm linked to Hillhead, however there is also an Hillhead Cottage that would plausibly be the correct link. All three of these links were found with the fuzzy matching. Both of these were found using matchkey Group E.

For the five incorrect links:

- Four were due to incorrect postcodes being recorded for the address in the development dataset. The links were then found using matchkey Group E. For example 20 Appletree Grove linked to 20 Appletree Lane, as the postcode recorded for Appletree Grove in the development dataset was actually the postcode for Appletree Lane.
- The final incorrect link was another erroneous link from matchkey Group E where the postcode appears to be correct, but some room information ends up linking to building information. In others the address in the development data included 'Room 1' and this linked to the building numbered 1 in that postcode.

While the results above from the random sample of 1,000 links gives an indication of how accurate the linking process is overall, it is dominated by links made using matchkeys in Group A with the minimal standardised variables.

In order to assess how accurate each group of links were an additional sample of 200 links from every other group was taken to increase the number of links reviewed. The results from the clerical review of these links, along with those from the sample of 1,000 above are shown in table 3, broken down by matchkey group.

Table 3: Summary of clerically reviewed links by matchkey group

Matchkey Group	Number of links reviewed	Correct link		Possibly correct link		Incorrect link	
		N	%	N	%	N	%
A - minimally standardised addresses	774	774	100.0	0	0.0	0	0.0
B - minimally standardised addresses	216	216	100.0	0	0.0	0	0.0
A - fully standardised addresses	276	276	100.0	0	0.0	0	0.0
B - fully standardised addresses	203	203	100.0	0	0.0	0	0.0
C - minimally standardised addresses	204	202	99.0	0	0.0	2	1.0
D- minimally standardised addresses	202	202	100.0	0	0.0	0	0.0
C - fully standardised addresses	226	223	98.7	0	0.0	3	1.3
D- fully standardised addresses	200	199	99.5	0	0.0	1	0.5
E	288	271	94.1	7	2.4	10	3.5
Fuzzy match	211	193	91.5	14	6.6	4	1.9
Total	2800						

For Groups C and D the two incorrect links are all for addresses where the same street name exists in two locations in or around Glasgow, and the only link made is to the location that seems least likely based on the postcode recorded. For example, there is a Broompark Drive in Glasgow, but also in Newton Mearns. In the Scottish Address Directory the Newton Mearns address also includes Glasgow as part of the address. Although the address being linked appears to be for the Broompark Drive in Glasgow, there is a flat number missing so no link is made there. However a link is

found to the property with the same number in Newton Mearns resulting in an incorrect link.

Group E contains the highest proportion of links that are clearly incorrect. Nearly all of these were due to incorrect postcodes being recorded in the development dataset. As the CCS addresses have been selected because of the postcode they are in and the value is hardcoded into the data, this should be far less of an issue with CCS and census addresses.

The fuzzy matching provides the lowest proportion of links that are clearly correct. The 6.6 per cent of addresses where the link was possibly correct were all instances where the address was for a farm or cottage with a certain name, but there are other addresses that could be plausible links as well despite not scoring well enough for the link to be recorded. The incorrect links occur when either:

- the address does not appear to be included in that postcode on the SAD, but another address is sufficiently similar to be recorded as a link.
- the address does appear in the SAD, but there is sufficient difference that another address is still linked ahead of it

Unmatched addresses

13.0 per cent of addresses in the development dataset (473,676 records) could not be linked to an address in the SAD. Some of the reasons for this are covered below, however it is expected that the higher quality of the address data in the CCS and the fact that the CCS is determined by a valid postcode will mean that this percentage is lower for CCS to census linkage.

Invalid postcodes: 113,775 addresses in the development dataset either have a missing postcode, or a postcode that does not appear against any addresses in the SAD. For these addresses, a link is only possible with matchkeys from Group C or D as all others require postcodes to be equal. 78,268 of these addresses are not linked, accounting for 16.5 per cent of the addresses that are not linked. This will be less of an issue when linking CCS addresses to census addresses as the vast

majority of addresses will all have valid postcodes due to the validation process in place.

For the addresses with a valid postcode, a random sample of 100 was taken to give some information on why a link was not found:

- For 33 addresses in the sample, there was insufficient information to make a link. This was usually due to missing flat information in the development dataset. For example the address in the development dataset was 1 Dalbeth Road, but within 1 Dalbeth Road there are flats so it is impossible to know which flat is being referred to.
- For 20 addresses in the sample, the address was for a flat but different numbering conventions had been used in the development dataset and the SAD. The conventions were sufficiently different that without knowledge of the specific building then it is not possible to determine which flat it is.
- For 18 addresses in the sample, the property does not appear to be included on the dataset. This could be because it is not included in the SAD, because it is listed as non-residential, or because it is numbered rather than named or vice versa.
- For 11 addresses in the sample, the address was a flat/room in a student halls of residence. However the SAD only includes the student residence as a whole, rather than including specific rooms or flats.
- For 10 addresses in the sample, although the postcode was a valid Scottish postcode, it was incorrect for the address. Ideally these would be identified by matchkey groups C or D, however differences in how the addresses were recorded prevented this. This is caused by spelling errors in the address and other differences in how the address is recorded such as differences in how flats are numbered or the locality and town information included.
- The remaining 8 all had quirks in how the address had been recorded that meant that the correct link was missed, however the correct link is obvious when manually searching for it. It may be possible to identify some of these links with some small modifications to the process.

Many of those cases above should not arise in the Census to CCS address matching exercise, although this will be dependent on the quality of enumerators' address records. The quality of recording will be a determining factor for the scale of clerical review required.

7. Strengths and Limitations

Strengths

The method has been developed in-house in SAS, a commonly used language within NRS. This means that it can be easy to adapt if necessary. This could be beneficial if addresses in the CCS are recorded in an unusual way that causes the matchkeys to be less effective than expected.

The process is quick to run, taking approximately 1.5 hours to attempt to link the 3.6 million addresses in the development dataset. For the CCS to Census linkage, the number of matchkeys for the CCS will be slightly higher than what could be produced from the development dataset. However as the CCS is only a sample of 50,000 addresses the time taken to run this process will be reduced.

The method has relatively high accuracy with approximately 99 per cent of links found being clearly correct. Some of the remaining 1 per cent are also likely to be correct. Any incorrect links should be identifiable when combined with the results from Census–CCS person linkage as this will highlight any links that suggest people are in different addresses in the census and CCS and require clerical review.

Limitations

There are some addressing issues that this method does not and cannot solve. For example, if the CCS has 1F1 13 High Street but the flats at 13 High Street are numbered 13/1, 13/2 etc. in the census data, then the link will not be identified.

8. Conclusion

The address linking method presented in this paper provides a more robust matching process than what was used in the 2011 Census. The overall success of the linkage method will depend on the quality of the addresses recorded in the CCS, but it would be expected that it will be higher than the 86.0 per cent linked from development dataset.

The quality of the links found with this method is high, with approximately 99 per cent being correct links. As well as having a high accuracy rate it is also quick to run.

It is difficult to estimate the scale of clerical review that will be required due to differences between the development dataset and CCS data, however the quality of data should mean it is not substantial. Also the CCS is a much smaller dataset than the development dataset used in this methodology. This is important in order to achieve the release of first census outputs within a year of collection.

Annex 1: Matchkey details

Please note the addresses used in this annex are fictitious and will have been modified to illustrate the selected examples.

Group A: at least property, street and postcode match exactly

Rationale: The addresses linked using the matchkeys from Group A have the following qualities:

- Postcode must be equal
- Both addresses have the information held within the fields for property names/numbers and street.

The matchkeys include the different combinations of including the other information. The reason for this is that in some cases information from the street variable in the CCS may be found in locality/town in the census dataset or vice versa.

A link is recorded if any of the CCS matchkeys are equal to any of the census matchkeys, i.e. a link will be recorded if CCS_A1 = CEN_A7. This may result in the same link being recorded multiple times if multiple combinations of matchkeys are equal, however this is resolved at the end of the process.

CCS: Concatenations of the variables below:

MK	Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
CCS_A1	Y	Y	N	N	N	Y
CCS_A2	Y	Y	Y	N	N	Y
CCS_A3	Y	Y	Y	N	Y	Y
CCS_A4	Y	Y	Y	Y	N	Y
CCS_A5	Y	Y	Y	Y	Y	Y
CCS_A6[11]	N	Y	N	N	N	Y
CCS_A7	N	Y	Y	N	N	Y
CCS_A8	N	Y	Y	N	Y	Y
CCS_A9	N	Y	Y	Y	N	Y
CCS_A10	N	Y	Y	Y	Y	Y

[11] Only for addresses where house number/name does not solely consist of numbers or flat information.

Census: Concatenations of the variables below:

MK	Organisation	Property	Building Number	Street	Locality	Town	Postcode
CEN_A1	Y	Y	Y	N	N	N	Y
CEN_A2	Y	Y	Y	Y	N	N	Y
CEN_A3	Y	Y	Y	Y	N	Y	Y
CEN_A4	Y	Y	Y	Y	Y	N	Y
CEN_A5	Y	Y	Y	Y	Y	Y	Y
CEN_A6	N	Y	Y	N	N	N	Y
CEN_A7	N	Y	Y	Y	N	N	Y
CEN_A8	N	Y	Y	Y	N	Y	Y
CEN_A9	N	Y	Y	Y	Y	N	Y
CEN_A10	N	Y	Y	Y	Y	Y	Y

Example: This group provides the most straightforward links. This example shows where the link is found as CCS_MK_A1 is equal to CEN_MK_A1 and would have been made even without cleaning the data. However, even in this straightforward example there are differences in how the address was recorded with the census including the locality information.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	FLAT 7	65 SKINNER STREET		DUMFRIES	DG21 7DX

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	FLAT 7	65	SKINNER STREET	LOCHRUTTEN	DUMFRIES	DG21 7DX

Group B: at least property information and postcode match exactly (spaces removed)

Rationale: When addresses are recorded, spaces can be added or removed from words erroneously depending on who has completed the information. For example HILLSIDE FARM could become HILL SIDE FARM. Matchkey Group B resolves instances like this by taking the matchkeys from Group A and removing the spaces. This results in the matchkey being one long string rather than a string of separate words. However, prior to removing spaces, underscores are inserted between any two numbers separated by a space to prevent the potential for confusion after concatenation. For example, 13/2 115 MAIN STREET would become 13/2_115MAINSTREET rather than 13/2115MAINSTREET.

This also resolves issues where one address had dashes between words (which were replaced with a space during cleaning) but the other did not separate the words at all, for example, TIGH-NA-MARA compared to TIGHNAMARA.

CCS: Concatenations of the variables below:

MK	Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
CCS_B1	Y	Y	N	N	N	Y
CCS_B2	Y	Y	Y	N	N	Y
CCS_B3	Y	Y	Y	N	Y	Y
CCS_B4	Y	Y	Y	Y	N	Y
CCS_B5	Y	Y	Y	Y	Y	Y
CCS_B6[12]	N	Y	N	N	N	Y
CCS_B7	N	Y	Y	N	N	Y
CCS_B8	N	Y	Y	N	Y	Y
CCS_B9	N	Y	Y	Y	N	Y
CCS_B10	N	Y	Y	Y	Y	Y

[12] Only for addresses where house number/name does not solely consist of numbers or flat information.

Census: Concatenations of the variables below:

MK	Organisation	Property	Building Number	Street	Locality	Town	Postcode
CEN_B1	Y	Y	Y	N	N	N	Y
CEN_B2	Y	Y	Y	Y	N	N	Y
CEN_B3	Y	Y	Y	Y	N	Y	Y
CEN_B4	Y	Y	Y	Y	Y	N	Y
CEN_B5	Y	Y	Y	Y	Y	Y	Y
CEN_B6	N	Y	Y	N	N	N	Y
CEN_B7	N	Y	Y	Y	N	N	Y
CEN_B8	N	Y	Y	Y	N	Y	Y
CEN_B9	N	Y	Y	Y	Y	N	Y
CEN_B10	N	Y	Y	Y	Y	Y	Y

Example: The links found in this group are generally where an extra space has been used in one of the addresses. In this case the word Springvalley has been split into two in the CCS, but is otherwise straightforward.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	1F3	35 SPRING VALLEY GARDENS		EDINBURGH	EH10 4FQ

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	1F3	35	SPRINGVALLEY GARDENS		EDINBURGH	EH10 4FQ

Both of these addresses would have the matchkey:
'1F3_35SPRINGVALLEYGARDENSEH104FQ'

Group C: at least property, street and town/locality information match exactly but non-matching postcodes

Rationale: Incorrect postcodes should not be an issue, or at least a very small issue, because of the way the address frames for the CCS and census are produced. For the CCS, addresses are included because they are in specific postcodes and hard-coded into the data so input error will be minimised. For census addresses there is validation on addresses before they are sent a questionnaire. However, some errors may occur when a respondent has corrected the address on a questionnaire although the number of cases where someone has corrected their address and made an error with the postcode should be small.

Despite this, Group C is included as a safety measure to check for links where postcode is not the same. The addresses linked using the matchkeys from Group C have the following qualities:

- The property, street and town in the CCS address must also appear in the census address.
- Postcode is ignored.

CCS: Concatenations of the variables below:

MK	Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
CCS_C1	Y	Y	Y	Y	Y	N
CCS_C2	Y	Y	Y	N	Y	N
CCS_C3	N	Y	Y	Y	Y	N
CCS_C4	N	Y	Y	N	Y	N

Census: Concatenations of the variables below:

MK	Organisation	Property	Building Number	Street	Locality	Town	Postcode
CEN_C1	Y	Y	Y	Y	Y	Y	N
CEN_C2	Y	Y	Y	Y	N	Y	N
CEN_C3	Y	Y	Y	Y	Y	N	N
CEN_C4	N	Y	Y	Y	Y	Y	N
CEN_C5	N	Y	Y	Y	N	Y	N
CEN_C6	N	Y	Y	Y	Y	N	N

Example: In this case the addresses clearly match apart from the postcode. In this instance the link is only found after cleaning as in the CCS record 26-1 is changed to 26/1 and HERMIT'S changed to HERMITS as part of the cleaning process.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	26-1	HERMIT'S CROFT		EDINBURGH	EH8 9RG

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	26/1		HERMITS CROFT		EDINBURGH	EH8 9RH

Both of these addresses would have the matchkey '26/1 HERMITS CROFT EDINBURGH'

Group D: at least property, street and town/locality information match exactly but non-matching postcodes (spaces removed)

Rationale: As with Group B, Group D modifies the matchkeys from the previous group slightly by inserting underscores between numbers separated by spaces and then removing all spaces. This allows extra links to be found that you would expect to have been found in Group C, but a mismatch in how spaces have been used has meant the link was missed in Group C.

CCS: Concatenations of the variables below:

MK	Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
CCS_D1	Y	Y	Y	Y	Y	N
CCS_D2	Y	Y	Y	N	Y	N
CCS_D3	N	Y	Y	Y	Y	N
CCS_D4	N	Y	Y	N	Y	N

Census: Concatenations of the variables below:

MK	Organisation	Property	Building Number	Street	Locality	Town	Postcode
CEN_D1	Y	Y	Y	Y	Y	Y	N
CEN_D2	Y	Y	Y	Y	N	Y	N
CEN_D3	Y	Y	Y	Y	Y	N	N
CEN_D4	N	Y	Y	Y	Y	Y	N
CEN_D5	N	Y	Y	Y	N	Y	N
CEN_D6	N	Y	Y	Y	Y	N	N

Example: In this case the addresses clearly match apart from the postcode. The initial cleaning of the addresses would correct HALLIDAY DR to HALLIDAY DRIVE in the CCS record. However, the fact that HOLMPARK is split into two words in the census is why the removal of spaces is required to find this link.

The fact that there is Organisation information in the census record that does not match the CCS information means that CCS_MK_D1 does not equal CEN_MK_D1 but instead this match would be found as CCS_MK_D1 equals CCS_MK_D4.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
HOLMPARK NURSING HOME	10	HALLIDAY DR	RUTHERGLEN	GLASGOW	G35 6DB

Census Address

Organisation	Property	Building Number	Street	Locality	Town	Postcode
HOLM PARK CARE HOMES LTD	HOLM PARK NURSING HOME	10	HALLIDAY DRIVE	RUTHERGLEN	GLASGOW	G53 3DB

Both of these addresses would have the matchkey 'HOLMPARKNURSINGHOME10HALLIDAYDRIVERUTHERGLENGLASGOW'

Group E: match within postcode using property numbers in the address only

Rationale: In many cases, a strong link can be found by only identifying flat and building numbers as well as the postcode. The three matchkeys in this group aim to take advantage of this. To do this all matchkeys in this group look through each word in the full address string. If the word contains a number, or has been cleaned to have a code that signifies flat information at the cleaning stage, for example G/F for a ground floor flat, then the word is included in the matchkey, otherwise the word is ignored. In addition to this, if the string 'FLAT X ' is part of the address, where X could be any letter, the flat letter is appended to the end of the first number in the matchkey. Some examples are provided below as an illustration:

Address	Matchkey
3/2 HIGH STREET GLASGOW G1 1TA	3/2 G1 1TA
GROUND FLOOR FLAT 16 DUNDEE STREET EDINBURGH EH6 6ST	G/F 16 EH6 6ST
FLAT B 14 UNION STREET AB1 2DE	14B AB1 2DE

Another advantage of the matchkeys in this group is that as they only consider the numbers and coded flat information from an address, typos in street names or towns are not an issue.

Matchkeys

MK	Description
E1	As described above
E2	For CCS records: If the E1 matchkey has more than 3 words (i.e. there are at least two parts that are not the postcode) then the first two words are swapped. So in the second example above this matchkey would become 16 G/F EH6 6ST. The reason for this is that on inspection of the addresses used during development, it was not unusual for the address to be written 16 GROUND FLOOR FLAT as an alternative to GROUND FLOOR FLAT 16. For census records: The E2 matchkey remains the same as the E1 matchkey, as if both CCS and CEN matchkeys had the orders swapped then you would not identify additional links.

E3	In addition to the steps above, the forward slashes are replaced with a space. So in the first of the examples above the matchkey would become 3 2 G1 1TA. This allows a match to be made if the CCS has the address in the 3/2 format but the census has FLAT 3 2 HIGH STREET
E4	In addition to the above, if there is a slash, the order of the numbers is reversed. Unlike for E2, this reversal occurs for CCS and census records as the intention of this matchkey is to link addresses where the flat numbering convention includes a forward slash to one where no slash was used. So in the first of the examples above the matchkey would become 2 3 G1 1TA. This allows a match to be made if the CCS has the address in the 3/2 format but the census has FLAT 2 3 HIGH STREET

Unlike the other matchkey groups, when making the comparisons for this group we only check CCS_E1 against CEN_E1, CCS_E2 against CEN_E2, CCS_E3 against CEN_E3 and CCS_E4 against CEN_E4 rather than doing every combination.

Example 1: In this example matchkey E1 finds the link which has been missed until this point due to a typo in the street name in the CCS. For both of these addresses matchkey E1 would equal 0/2 998 G31 4HG.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	0/2	998 SPRINGIELD RD		GLASGOW	G31 4HG

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	0/2	998	SPRINGFIELD ROAD		GLASGOW	G31 4HG

Example 2: In this example matchkey E1 would equal '14H PH1 2TN' for both addresses. This example also shows how differences in street name do not prevent a link being made.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	14H	LESLIE COURT		PERTH	PH1 2TB

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	FLAT H	14	LESLIE COURT FAIRFIELD AVENUE		PERTH	PH1 2TB

Example 3: When developing this code it was noticed that in some datasets the flat numbers were being recorded after building numbers. This situation should be less common in the CCS as the type of information in each field is defined and CCS interviewers will have guidance to help with this. In this example matchkey E2 is '1F1 34 EH11 2LG' for both addresses as the first two components of the matchkey are switched in the CCS.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	34	MURIESTON CRESCENT	1F1	EDINBURGH	EH11 2LG

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	1F1	34	MURIESTON CRESCENT		EDINBURGH	G31 4HG

Example 4: In this example matchkey E3 is equal to '8 88 EH14 5SD'. In many instances 8/88 would mean Flat 88, 8 GREAT NORTHERN ROAD, however when inspecting this example more closely there is no such address so it seems clear that this must be a correct link. If there had also been a Flat 88 8 GREAT NORTHERN ROAD then that link would be picked up with matchkey E4 and both possibilities would be recorded.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	8/88	GREAT NORTHERN ROAD		EDINBURGH	EH14 5SD

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	FLAT 8	88	GREAT NORTHERN ROAD		EDINBURGH	EH14 5SD

Example 5: In this example matchkey E4 is equal to '6 1 G83 2HB'. As mentioned in the previous example, if there was a Flat 1, 6 WILKIE WALK in this postcode then it would have been recorded from matchkey E3 as well.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	1/6	WILKIE WALK		DUMBARTON	G83 2HB

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	FLAT 6	1	WILKIE WALK		DUMBARTON	G83 2HB

Link using fuzzy matching on house names within the same postcode

Rationale: This Group finds matches where there are other small mismatches in how a property has been recorded, for example, Hillside Cottage rather than Hillside Farm Cottage. It also captures occasions where a minor typo has been made in an address.

Unlike the other groups, this group does not create a matchkey and instead does a comparison between the CCS address and all census addresses in the same postcode. This comparison is made using a string comparison algorithm which was developed as part of the name linking process. This algorithm produces two scores that measure the similarity between two strings. The first of these scores is based on the longest string of consecutive characters that the two strings have in common. The second is based on the number of substitutions, deletions, insertions, transpositions and jumps required to convert one string into the other. All links found where the scores are below a certain threshold are then recorded. The thresholds were chosen based on a clerical review so that the majority of links recorded are correct links.

The variables used to make this comparison are:

CCS:

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
Y	Y	N	N	N	N

Census:

Organisation	Property	Building Number	Street	Locality	Town	Postcode
Y	Y	Y	N	N	N	N

Example 1: Finding a link despite a minor typo: In this example there is a slight difference in the house name, meaning the link has not been made previously, but it is clearly the same address.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	LOCK HOUSE	STROMNESS		ORKNEY	KW17 3BU

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	LOCH HOUSE			STROMNESS	ORKNEY	KW17 3BU

Example 2: A farm where the word Farm has been included in the CCS but not in the census. This is a common difference where the address for a farm does not always specify that it is a farm, but when the addresses in that postcode are inspected it is clear they are the same address.

Establishment name	House Number/Name	Street	Addressline3	Town	Postcode
	WOODSIDE FM	CURRIE		EDINBURGH	EH14 5ST

Organisation	Property	Building Number	Street	Locality	Town	Postcode
	WOODSIDE			CURRIE	EDINBURGH	EH14 5ST

Annex 2: CCS records where an alternative address should be used

In a small number of instances, a household in the CCS will have moved address since the census took place. In this situation the respondent is asked what their address was on census day and this address is linked to the census records rather than the address the respondent is at during the CCS. However, the response to this question is not broken down into House name/number, street, addressline3, town and postcode. Instead the address is recorded in just two variables, one for the address and one for the postcode. This means that the full range of matchkeys used for the majority of CCS addresses cannot be created. Instead alternative matchkeys that are similar to those in each group are created. As the format of the address limits the matchkey variations that can be produced there may be a reduced linkage rate for these addresses. In any case, the number of CCS addresses where this situation arises will be relatively small so it should not be a large burden if clerical review is required in this situation.

These are described below:

Group A

A1: The full address and postcode

A2: The address until the final appearance of a 'street' signifier[13] and postcode. This acts as an approximation of only looking for property name/number and street, although it will not work in cases where no street is included or there are typos.

Group B

As with the main set of matchkeys, Group B takes the Group A matchkeys and removes spaces after inserting underscores between any consecutive numbers separated by a space.

Group C

C1: The full address without postcode

[13] These signifiers are the words Avenue, Brae, Court, Crescent, Drive, Gardens, Grove, Lane, Loan, Parade, Park, Place, Quadrant, Road, Rise, Square, Street and Terrace.

Group D

As with the main set of matchkeys, Group D takes the Group C matchkeys and removes spaces after inserting underscores between any consecutive numbers separated by a space.

Group E

The matchkeys for this group can be derived in the same manner as for the main set of addresses.

Fuzzy Matching

There is no obvious way to isolate the house name when the address is provided in one string. Instead the whole string could be used instead, although it is less likely to identify a link.

Annex 3: Glossary

Term	Definition
False link	A link that has been made, but is for two different addresses
Fuzzy matching	Fuzzy matching is a technique used to find instances where two strings of text are believed to match despite not being identical.
Link	Two addresses that have been connected
Matchkey	Strings that are a sub-string of the full address and are used to make comparisons between two addresses.

Annex 4: Information Governance

The 2016 Health Activity dataset used in the development and testing of this method of address matching was provided to the NRS Admin Data team as part of a project to produce household estimates from administrative data. Part of this project required a Unique Property Reference Number (UPRN) to be added to the data by linking the address to the Scottish Address Directory to attach a UPRN to records where possible. As a by-product of this project the findings have then been adapted to create the method can be used for CCS to Census linkage.

More information on this administrative based population and household estimates can be found:
Administrative Based Population and Household Estimates (DPIA)