

Scotland's Census 2022

**Overview of  
Edit and Imputation  
for  
Scotland's Census 2022**

September 2020

## Contents

1. Plain English summary/Abstract .....	4
2. Executive summary/Abstract.....	5
3. Introduction and Background .....	6
3.1 Prevention of errors .....	8
3.2 Choice of imputation methodology & software .....	10
3.3 CANCEIS implementation of imputation methodology .....	12
3.4 Methodological research for process design .....	15
4. Summary of Methodology used in 2011 .....	16
4.1 Modularisation .....	16
4.2 Processing Units .....	19
4.3 Partial codes .....	20
4.4 Relationships .....	22
4.5 Edit Rules .....	25
4.6 Administrative Data .....	26
4.7 Voluntary Questions.....	27
4.8 Ad-hoc data adjustments .....	28
4.9 Quality Assurance .....	29
4.10 Audit and Metrics.....	29
5. Proposed Methodology for 2022.....	30
5.1 Modularisation .....	31
5.2 Processing Units .....	33
5.3 Partial codes .....	34
5.4 Relationships .....	36
5.5 Edit rules .....	40
5.6 Administrative Data .....	42
5.7 Voluntary Questions.....	45
5.8 Ad-hoc data adjustments .....	46
5.9 Quality Assurance .....	46
5.10 Audit and Metrics .....	48
6. Conclusions .....	49
7. References .....	51
8. Annex of supporting information .....	52
8.1 Definitions .....	52
8.2 "Similarity" in CANCEIS and the use of predictors .....	54

8.3	Relationship Algorithm 1: Common errors to fix.....	55
8.4	Edit Rules in 2011 .....	56
8.4.1	Household .....	56
8.4.2	Demographics .....	56
8.4.3	Culture .....	58
8.4.4	Health.....	58
8.4.5	Labour Market .....	58
8.5	Processing Units in 2011 .....	59
8.6	Item non-response rates .....	59

## 1. Plain English summary/Abstract

Scotland's Census 2022 asks every person in the country questions about themselves and the people they live with. Respondents are told that they must answer every question, except for those labelled as voluntary, and any questions which the respondent is instructed to skip if they are not relevant (this is an automatic process for online respondents). For example, the question about marital status will not be asked of people under the age of 16.

Despite every effort to help and encourage respondents to fill out the questionnaire as accurately and completely as they can, there will inevitably be returns which are missing answers to some questions. There will also be some respondents who make mistakes when answering questions, which can lead to inconsistencies across a response. For example, if someone writes the current year instead of their birth year into the Date of Birth field, they will appear to be 0 years old and yet may be married, have qualifications, a job, and so on.

The Edit and Imputation process is about identifying these missing and inconsistent responses, and filling in the blanks and correcting the inconsistencies using robust statistical methods to produce plausible results. This process is not a crystal ball: we cannot say for certain that the value we assign is true for an individual, but the overall effect will be that the outputs produced from census data will accurately reflect the population.

The main method we use for Edit and Imputation is called donor imputation. For each record which needs to be fixed, we look for similar records in the census dataset and then we copy-and-paste responses from the donor record, in order to fill in the blanks or correct inconsistent responses.

In this paper we summarise the Edit and Imputation methodology used in Scotland's Census 2011, and outline the main improvements which we have made for Scotland's Census 2022. These improvements will enhance the quality of imputation,

and in some cases decrease the time it takes to run the data through the Edit and Imputation process.

## 2. Executive summary/Abstract

Scotland's Census 2022 is a self-completed questionnaire which asks a considerable number of questions about every person usually resident in the country. With the exception of voluntary questions and questionnaire routing, every question is mandatory.

Despite every effort to help and encourage respondents to fill out the questionnaire as accurately and completely as they can, there will inevitably be a certain level of non-response to each question, and there will be a certain level of respondent error leading to invalid values and inconsistencies within a response.

The Edit and Imputation process is about detecting and correcting these missing and inconsistent responses, using robust statistical methods to produce plausible results. This process will generate some error, but the methodology ensures that the imputed dataset accurately reflects the population distributions and cross-distributions for all variables.

The main method we use for Edit and Imputation is called donor imputation. For each record which needs to be fixed, we select similar records in the census dataset and then we impute responses from the donor record, in order to resolve the missing, invalid or inconsistent responses.

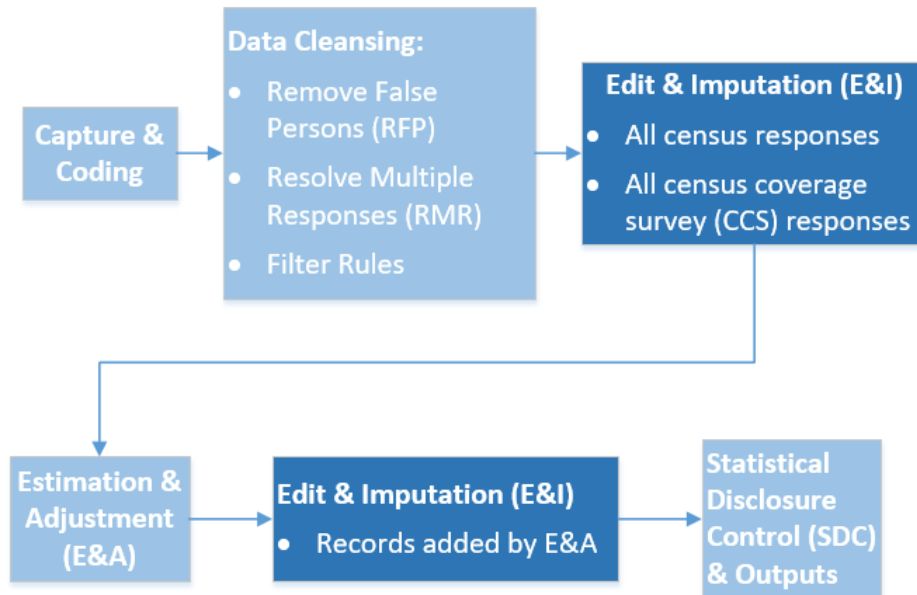
In this paper we summarise the Edit and Imputation methodology used in Scotland's Census 2011, and outline the main improvements which we have made for Scotland's Census 2022. These improvements will enhance the quality of imputation, and in some cases decrease the time it takes to run the data through the process.

### 3. Introduction and Background

Although every effort is made to collect full and accurate census responses, inevitably there will be some incomplete returns due to respondents not answering all mandatory questions. Census returns may also contain invalid and inconsistent responses, which may be due to respondent error or procedural error such as limitations of the data capture process for paper questionnaires.

There is an expectation that the main census outputs are complete and consistent, so this is dealt with by Edit and Imputation during data processing, before the production of census outputs (Figure 1). With unrestricted record-level access to the complete dataset, we can make the best use of all the information available to accurately impute the dataset using statistically robust methods.

Not all data users want to work on imputed data, for example some researchers who use record-level (“microdata”) extracts may want to use unimputed data and use complete case analysis (where records with missing values of interest are ignored) or other methods of imputation. An example of this is the Scottish Longitudinal Study, which links census data to other data sources to track individuals over time: an imputed value may result in an individual appearing to have temporarily changed their marital status, just as a result of imputation. Thus we will produce imputation flags, which indicate which variables have been imputed in each record (see Sections 4.10 and 5.10).



**Figure 1: Where Edit and Imputation fits into Statistical Data Processing**

Item-level Edit and Imputation is the detection and correction of missing, invalid and inconsistent values in census responses. This is applied to the dataset after it has passed through Data Cleansing but before Estimation and Adjustment, as shown in Figure 1.

Item-level edit and imputation is also applied to the dataset after Estimation and Adjustment, in order to complete the records which have been added by coverage adjustment. These added records are copied from donors in the adjustment process, so to avoid outright duplication, only a small selection of key variables (such as age, sex, ethnicity) are copied into these new records. The rest of the information is imputed separately using a second edit and imputation process as shown in Figure 1.

Matching processes, such as the Census-to-Census Coverage Survey matching which feeds into Estimation and Adjustment, is performed on unimputed data. It is important that the records are correct at an individual level to ensure accurate matches.

Note that item-level imputation is different to unit-level imputation, which is the insertion of synthetic records to deal with people and households missed by the census - this is part of the process known as Estimation and Adjustment [1].

All variables used in census outputs are either subject to Edit and Imputation, or are derived from variables subject to Edit and Imputation, with the exception of voluntary questions, where non-response is acceptable.

### 3.1 Prevention of errors

Responses are **missing** if the respondent skips a mandatory question without being instructed to do so. For example in 2011 the “long-term health conditions” question had a high rate of non-response - based on their responses to other health questions, it was assumed that this was because respondents thought the question was not relevant to them, and did not see the tick-box at the end of the question for “no condition” (see Section 4.8)

A value may be **invalid**, for example, if it is out of range (e.g. 612 years old), if more than one box was ticked for a single-tick question (e.g. you cannot answer “yes” and “no” to the question on full-time education), or if the value in a text field does not yield a valid response (e.g. “pregnant” is not a health condition lasting, or expected to last, longer than 12 months).

A value is **inconsistent** if it contradicts other information. For example, if an individual is aged 7 and is listed as a parent of another household member, the age and the relationship will both be flagged as inconsistent, and one of these values must be changed in order to resolve the inconsistency.

In 2011 about 20% of responses were online and the rest were on paper. In the next census, for the first time, the majority of people will be encouraged to complete their questionnaire online.

We can make use of this technology to reduce the number of errors. For example, we can limit acceptable values, remind respondents to provide a response if they try



to skip a mandatory question without answering it, and provide opportunities to review responses. We can enforce “select one response only” type questions with radio buttons. Our type-ahead lists on text fields will reduce the number of incorrectly spelled responses for questions such as long-term conditions and occupation and industry. Likewise we are using address lookups to validate addresses for place of work/study and address one year ago.

Online responses will naturally not contain some of the errors we see in paper responses, such as scored-out questions being automatically scanned as ticks, and handwriting recognition issues which are particularly problematic for numeric fields.

When answering the date of birth question, the calculated age will be displayed: this should reduce the number of people accidentally giving us the current year instead of their birth year. An error message will be displayed if an answer is out of bounds, for example if the date of birth entered is after census day.

The relationship question will be much easier to answer online because we can use piped names and drop-down menus to display each relationship as “Alex is the [RELATIONSHIP] of Bob” instead of a matrix of tick-boxes seen in Figure 2. Respondents will also be able to see each reciprocal relationship as they work through the relationships question: if Alex is the parent of Bob, then on the page of Bob’s relationships, they will see that Bob is the child of Alex. At the end of the questionnaire there will also be a review page, for respondents to check their answers before submission.

Respondents who attempt to skip a question without providing a response will be shown a prompt reminding them to provide a response. It is possible in most cases to continue through the questionnaire without answering the question, with the option of going back to the question later.

We are aware that there will be a demographic difference between online respondents and paper respondents, with many older people and other digitally excluded demographics preferring to respond by paper or being unable to respond

online. Therefore response mode will be taken into account as part of the edit and imputation strategy.

### 3.2 Choice of imputation methodology & software

The census dataset contains a large number of categorical and numeric variables, with complex interactions between variables and even between records (individuals within a household). The inclusion of categorical variables, as well as the numerous potential between-variable and between-record inconsistencies, means that imputation using average values and other arithmetic solutions is not suitable for household census datasets.

As a self-completed survey of the entire population, data is often not missing completely at random, and the characteristics reflected in other given responses can influence the likelihood of non-response (this is known as “Missing Not At Random”, or “MNAR” for short). For example, a retired person may not the question about student/schoolchild status, as they consider it not relevant to them and “obvious” that they are not a student or schoolchild at their age. As a result, complete case analysis (in which only complete records are considered) may disproportionately exclude certain subsets of the population.

Prior to the methods discussed in this section, automatic imputation of census data relied on long, complex deterministic algorithms which would deduce a suitable value to impute based on the information given. The **Fellegi-Holt** method [2], published in 1976, is based on the principle that inconsistencies can be defined using a set of self-contained edit rules, which define what cannot occur in the data rather than what must be imputed (See Section 4.5 for more information). The system deduces the smallest number of changes required in order to resolve the edit rule failures.

The suggested basis for imputation with this method was hot-deck donor imputation. **Donor imputation** copies values from another “donor” record into the failed record. When the donor record is from the same dataset as the failed record, this is known as **hot-deck** imputation.

Advantages of the Fellegi-Holt method over previous methods include:

- By design, minimal changes were made to observed values
- Joint distributions maintained
- Changing one edit rule would not require extensive re-write of the system
- Easier to understand edit rules (described using a series of statements) rather than deterministic algorithm (described using a flowchart of possible combinations)
- Log of records failing edit rules can be used for later analysis of data and methods
- All available information used to inform imputation
- Adaptable to different surveys: bespoke programming not required for each new survey or dataset

Many international edit and imputation systems were subsequently created and applied to census data using this method, including CANEDIT and GEIS in Canada, DISCRETE and SPEER in the United States, CherryPi in the Netherlands, and EDIS in the United Kingdom.

Researchers at Statistics Canada built upon the Fellegi-Holt method by developing the **nearest-neighbour imputation methodology** [3] [4]. This method begins by searching for potential donor records similar to the one requiring imputation (so-called nearest neighbours), and then considers the smallest number of values to copy to the failed record in order to resolve the edit rule failures. The software which implements this method is called the Canadian Census Edit and Imputation System (CANCEIS) <sup>1</sup>.

Nearest-neighbour imputation as implemented in CANCEIS allows the simultaneous imputation of categorical, numeric and alphanumeric variables over large datasets, with a large number of user-defined edit rules. The imputation method itself is highly customisable, allowing tuning of the software to the dataset being imputed through a

---

<sup>1</sup> Contact [canceis@canada.ca](mailto:canceis@canada.ca) for more information on CANCEIS.

large number of system parameters. Processing speed is much improved over previous systems, suitable for large datasets such as censuses. By imputing variables simultaneously, joint distributions are preserved by reducing the likelihood of imputing two common values which never appear together in observed data. By searching for nearest neighbours, the donor is likely to be a member of the same subsets of the population to which the failed record belongs, resulting in more plausible imputation actions [4].

One limitation of this method of imputation is that it relies on a large pool of potential donors, to increase the likelihood of finding a donor record which closely resembles the record being imputed, and to reduce the likelihood one record being “cloned” too many times as a donor. This is why this method is suitable for a census, but may not be suitable for smaller surveys.

Another limitation is that it relies on an input of good quality data. Firstly, the ratio of failed records to potential donors must be reasonable for similar reasons outlined above. Secondly, a failed record should contain plenty of valid responses to other questions in order to find similar records to use as donors. For example, we would have a much better chance of accurately imputing someone’s main language if we know about their country of birth, national identity, and ethnicity. This quality of input data cannot be guaranteed for all records, but the use of public engagement to encourage full responses is vital to maximising overall item-level response rates.

### 3.3 CANCEIS implementation of imputation methodology

As mentioned in Section 3.2, CANCEIS uses nearest-neighbour hot-deck donor imputation methodology. There is an assumption in CANCEIS that geographically close records are demographically similar, and in fact the CANCEIS system assumes that the dataset is sorted geographically as best possible<sup>2</sup>, so that each

---

<sup>2</sup> It is not possible to perfectly map a two-dimensional map onto a one-dimensional list, but we can make use of space-filling curves to number local authority areas and planning areas. Planning areas are geographic partitions of local authority areas, primarily used for enumeration. Each planning area contains about 500 households. We can sort the dataset using these numbered geographic areas as well as sorting by postcode within each planning area. This ensures that each household is close to at least 500 households from the same immediate area, usually more.

record (household or individual) is close in the dataset to other geographically close records.

An example of this process is shown in Table 1 and Table 2, using a fake dataset in which record 3 is missing a value for occupation. This record fails the edit, since there is a missing value, and so it undergoes donor imputation.

ID	SEX	AGE	QUALS LEVEL	INDUSTRY	OCCUPATION	EDIT
1	M	21	1	Retail	Assistant	Pass
2	F	41	3	Transport	Pilot	Pass
3	F	24	1	Retail	<i>(missing)</i>	<b>Fail</b>
4	F	40	2	Retail	Store Manager	Pass

**Table 1: Example of a record (#3) which fails the edit (*not real data*)**

The most similar record to number 3 is record 1: the individual is of a similar age, with the same level of qualifications, and works in the same industry. Thus the value for occupation is copied from record 1 to record 3. With this new value, record 3 now passes the edit.

ID	SEX	AGE	QUALS LEVEL	INDUSTRY	OCCUPATION	EDIT
1	M	21	1	Retail	Assistant	Pass
2	F	41	3	Transport	Pilot	Pass
3	F	24	1	Retail	Assistant	Pass
4	F	40	2	Retail	Store Manager	Pass

**Table 2: Record 3 has been imputed using record 1 as a donor (*not real data*)**

The Edit and Imputation process in CANCEIS is summarised as follows:

1. **Edit:** All records are checked for missing, invalid and inconsistent values. Records marked as “pass” or “fail”.
2. **Imputation:** Each failed record is imputed as follows.

- a. **Donor search:** Failed record selected. Passed records which are similar to the failed record are selected as potential donors. Ranked by similarity, a shortlist is selected.
- b. **Potential imputation actions generated:** An imputation action is a selection of variables from a potential donor, which are copied-and-pasted to the failed record to fix it. From the shortlist of potential donors, all possible imputation actions are considered and a shortlist is created based on principles of minimal change to the failed record, and plausibility (similarity of the imputed record to the donor record).
- c. **Imputation action selected:** An imputation action is randomly selected from the shortlist. The higher quality imputation actions (similarity/plausibility score) are more likely to be selected. The selected values are copied-and-pasted from the donor to the failed record.

More information on how similarity is measured, how shortlists are generated, and how imputation actions are chosen, can be found in Section 8.2.

One important feature of this is the **donor pool**. This refers to all records which passed the edit, and can be used as potential donors. The larger this set is, the more likely it is that a donor record will be a very good match for a failed record, in terms of similarity. In particular, it increases the quality of imputation of outlier records as it is more likely that there is another record in the dataset which is also an outlier. We set a maximum number of times each record can be used as a donor, which helps to minimise large numbers of identical-looking records. The ratio of potential donors to failed records is therefore important in being able to find suitable donors, due to limits on donor reuse.

### 3.4 Methodological research for process design

The improvements outlined in this paper represent the culmination of four years of methodological research with the aim of improving data quality and processing efficiency.

This research has been primarily conducted using data from Scotland's Census 2011, as the closest proxy to how we expect the data to look in Scotland's Census 2022. There will, of course, be differences to the structure of the dataset, for example the inclusion of new questions and the redesign of some older questions. There will be differences to the demographics in the dataset, such as the distribution of ages, occupations, and household structures. There will also be differences in the quality of the data, as we expect that the majority of responses being online will improve overall quality of the data being fed in to the edit and imputation process.

We also have the data from the 2019 rehearsal, a small-scale, entirely voluntary survey based on the new design of the census questionnaire at the time. We are very grateful to all who participated in the 2019 rehearsal as the data collected helps us to refine our systems in the run-up to the census. The rehearsal dataset can provide us with some insight into how the new and changed questions are likely to be answered in the upcoming census, however we are aware that this survey is not representative of the population as a whole, and is skewed towards demographics who had the time and the inclination to fill in a very long survey entirely voluntarily. Our conclusions based on rehearsal responses therefore must be tempered accordingly.

The general method used for process design is as follows:

1. Start with a dataset which contains no missing, invalid or inconsistent values.
2. A random selection of records will have values removed for the variable(s) being tested.
3. The dataset will then be imputed, using the proposed method being tested, and separately using the old control method.

4. The results of the two imputation runs will then be compared to the original dataset, to see how close the imputed values are to the original values.

An example of this testing can be seen in Section 5.3, with the results of testing the new method for imputing occupation using partial codes.

#### **4. Summary of Methodology used in 2011**

Scotland's Census 2011 was the first census for which Scotland did its own data processing. All of the code used by National Records of Scotland (NRS) in 2011 had originally been developed by the Office for National Statistics (ONS) for the England and Wales census.

There was a considerable amount of work required to modify the code for Scottish data, in collaboration with colleagues at ONS and the CANCEIS support team at Statistics Canada. In particular, the relationship algorithms (Section 4.4) required extensive adaptation from the original ONS design, as there was one fewer household member on the main Scottish household form, and the Scottish continuation form asked about different relationships.

In addition to this, the software CANCEIS (Section 3.3) had never been used before by any UK census office, and required significant adaptation for use in the UK context. For example, as it was, CANCEIS could not impute partial postcodes as expected, and a workaround had to be developed (see Section 4.3).

Since the 2011 census NRS have been able to study the 2011 data and the CANCEIS software and improve the Edit and Imputation methodology for 2022.

##### **4.1 Modularisation**

A module can be thought of as a collection of variables which are imputed at the same time. By imputing variables at the same time rather than one after the other, we can resolve inconsistencies between variables based upon the characteristics of



the entire record, rather than by which variable is imputed first. Potential imputation actions are assessed on how well they minimise the change to the failed record, and how plausible they are in terms of similarity between the imputed record and the donor record. Details of this are in Section 8.2.

The variables are grouped together thematically so that they contain variables which are, to an extent, related to and predictive of each other. For example, a person's perception of their general health is related to whether they have any long-term health conditions, but it is not directly related to their skills in Gaelic and Scots.

Modularisation in 2011 was based on research conducted by the Office for National Statistics [5]. The household variables (cars, central heating, etc.) were imputed in one module. The person variables were grouped into four main modules to be imputed as follows:

<b>Demographics</b>	<b>Culture</b>	<b>Health</b>	<b>Labour market</b>
Age	Address 1 year ago	Carer	Qualifications
Sex	Country of birth	Disability	Ever worked
Marital status	Date arrived in UK	Health	Hours worked
Full-time student	Ethnicity	Long-term	Employee status
Term-time location	National identity	conditions	Supervisor
Relationships	Language questions		Industry
Economic activity			Occupation
			Work/study address
			Method of travel

**Table 3: Modularisation of non-voluntary person variables in 2011  
Showing in which module each variable was imputed**

We also included variables in a module which were not to be imputed, but which were used purely as predictors. For example, after age was imputed in Demographics, it was included as a predictor variable in the culture, health and labour market modules, as it was predictive of migration, health, qualifications & employment questions.

Many of the predictor variables listed in Table 4 were grouped into categories instead of, or as well as, including the specific values as predictors. For example, in the culture module, country of birth was a predictor both by specific country, but also as inside/outside the UK and by continent. This was a way of programming the software to understand that, say, Bolivia and Peru are much more similar countries than Bolivia and the UK for predicting culture variables.

<b>Demographics</b>	<b>Culture</b>	<b>Health</b>	<b>Labour market</b>
Marital status	Country of birth	Marital status	Marital Status
Country of birth (in/outside UK)	Ethnicity	Country of birth	Economic activity
Economic activity	Address 1 year ago	Economic activity	Full-time Student
Full-time student	Main language	Ethnicity	Ethnicity
Relationship to person 1	Date of arrival in UK	General health	Full-time student
Term-time location	National identity	Long-term health conditions	Industry
			Occupation
			Qualifications
Age			
Sex			
Enumeration location			
Response mode (paper/online)			
Hard-to-count code <sup>3</sup>			

**Table 4: Modularisation of non-voluntary person variables in 2011**  
**Showing the predictors in each module (and predictors for all modules)**

In 2011, for the four person modules, household records were edited and imputed as households rather than as individual persons. This means that the characteristics of the entire household were used to find similar donors. For example, if one person in a four-person household had a missing occupation, the characteristics of the other

<sup>3</sup> An index indicating how willing households within a Planning Area will be to respond to the census [13]

three household members were taken into account in trying to find a similar household to use as a donor.

Communal establishment records were imputed as individuals, but they were still processed through demographics, culture, health and labour market modules, with the main difference being that there were no relationships to impute, and there was a question on position in communal establishment which needed to be imputed.

In order to impute records as households in CANCEIS, the data had to be partitioned into households of the same size (same number of usual residents). This is because donor records needed to be the same size as the failed records they were helping to impute. These partitions were called **strata**, and the stratum number referred to the household size for that partition. For reasons explained in Section 4.2, only household persons 1-5 were imputed as households. Persons 6+ were imputed as individuals.

## 4.2 Processing Units

In 2011 the census data was partitioned into 10 geographically-based processing units - see Annex 8.5 for the list. These datasets were processed in isolation from each other, and remained the same from data collection right through to the production of output variables.

The main advantage of this for Edit and Imputation was that it required less processing power to deal with smaller datasets of around 500,000 individuals. The software at the time could be slow to run and computers were not as powerful as they are now.

A disadvantage of the use of processing units was that it reduced the number of larger-sized households being processed together. Thus when the processing unit was split into strata of households of the same size for imputing as households (see Section 4.1), there were not enough households of size 6+ to be able to run these through CANCEIS as households. Therefore the first five household members in

larger households were treated as a five-person household in stratification, while the remaining household members were imputed as individuals.

To illustrate this, suppose there were 100 households each containing 6 usual residents in Scotland<sup>4</sup>. There would be at best 10 households in each processing unit. A dataset of 10 households is not large enough to guarantee quality imputation, as there aren't enough potential donors. Additionally, the more people there are in a household, the more chances there are for error, so the more likely it is that a household will need to be imputed. Thus edit failure rates were expected to be higher in larger households, so the number of potential donors was insufficient to be able to impute these records as households.

The main downside of being unable to impute persons 6+ as members of their own household was that we needed an alternative way of imputing household relationships. There were too many records to be imputed manually, so an automated process was required. The solution was Relationship Algorithm 3 (see Section 4.4)

### 4.3 Partial codes

Some variables, such as postcode, industry and occupation, are encoded in a hierarchical manner. For example, a full postcode such as EH12 7TF provides a street-level location, but there are three higher levels of geography encoded within the postcode. The area (EH - Edinburgh area), the district (EH12 - a geographic subsection of the EH area), and the sector (EH12 7, a cluster of streets within EH12).

There were two questions involving postcodes in 2011: address 1 year ago and place of work/study. Particularly with the work/study question, respondents did not always know the work/school postcode, but could provide other lines of the address. The coding team tried to derive the postcode from the rest of the address, but where this was not possible, sometimes a partial postcode could be gleaned.

---

<sup>4</sup> Not real data. Number chosen for simplicity.

In 2011, use of partial codes was limited to postcodes. In Section 5.3 we explain how we have expanded this to industry and occupation codes (SIC and SOC).

The partial postcode could be used to impute a full postcode - however in 2011 we did not know how to do this in CANCEIS, so it was imputed using a deterministic algorithm, which worked as follows.

In the first stage, the data was split into two groups: those with complete and those with partial postcodes. The records with complete records functioned as donors for those with partial postcodes. In the second stage, the data was split into four distinct groups: primary school children, secondary school children, university students and non-students. Imputation of postcodes was carried out separately for each of these groups.

For schoolchildren, the most common full postcode in their output area, for their age group (primary or secondary), was considered first. Where there was a match between the most common full postcode and the partial postcode, the full postcode was used to resolve the partial postcode. Where no match was found, the most common postcode for the combination of business and occupation group for the output area was considered. (Persons aged below 16 do not have to supply this information so the value of this variable for this group was always 'no code required'.) If the partial postcode matched the most common full postcode for the combination of business and occupation group in the output area this postcode was used to resolve the partial postcode.

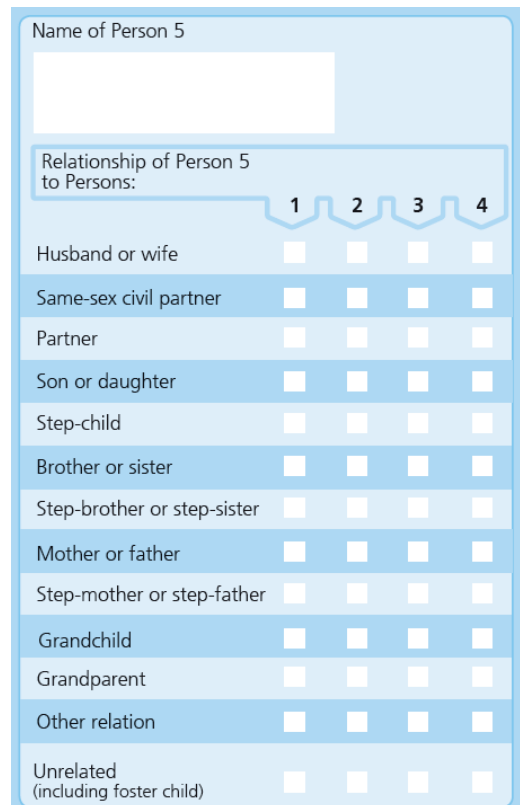
For non-students, the business and occupation group in the output area was concatenated and for each combination of business and occupation group the most common postcode was captured. Records with a partial postcode were matched on the combination of business and occupation group. If the most common postcode for the combination of business and occupation group matched the partial postcode provided by the respondent, this postcode was used to complete the partial record. If no match on business and occupation group was found, the most common full

postcode for the output area that matches the partial postcode was used to complete the partial postcode.

#### 4.4 Relationships

The relationships question was particularly tricky for some respondents to answer in 2011. Common errors included:

- Getting relationships the wrong way round
- Ticking the relationship to the person filling in the form rather than the relationship to the person indicated in the column
- Not filling in the persons in the same order consistently throughout the household form, so that age, sex and marital status did not align with the recorded relationships.



Name of Person 5

Relationship of Person 5 to Persons:

	1	2	3	4
Husband or wife	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Same-sex civil partner	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Partner	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Son or daughter	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Step-child	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Brother or sister	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Step-brother or step-sister	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mother or father	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Step-mother or step-father	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Grandchild	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Grandparent	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Other relation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Unrelated (including foster child)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 2: Relationships question in 2011 for household person 5

The imputation of relationships in 2011 was as follows:

1. Relationship Algorithm 1
2. Relationship Algorithm 2 - part of original design but ultimately not included in 2011 live processing
3. Donor imputation in CANCEIS in demographics module
4. Relationship Algorithm 3

**Relationship Algorithm 1** was a deterministic algorithm applied prior to donor imputation. It corrected the most common respondent errors, for example listing relationships the wrong way round (“I am his parent” instead of “He is my child”) and missing the more obvious relationships (three-generation relationships, missing siblings/parents).

Since this was a deterministic process, the algorithm used very strict criteria to identify records on which it would act. For example when looking at a parent-child relationship where the child is older than the parent, the following criteria must all be met before the parent-child relationship was reversed:

- Younger person is (mis)reported as parent of older person
- The age gap between the two people is at least 13 years
- The older person is at least 16 years old
- The younger person is no more than 30 years old
- The younger person has a marital status “single”
- The younger person has no partner, spouse or civil partner in household

Relationship Algorithm 1 significantly reduced the number of households requiring imputation in the demographics module, which in turn increased the quality of imputation and chances of successful imputation by donor.

The common errors fixed by Relationship Algorithm 1 are listed in Section 8.3. As this was a deterministic algorithm, the conditions required to apply a rule were quite strict. For example, in order to apply a rule involving siblings, the people to be made

siblings cannot have more than a 20 year age difference. This was not a general restriction on the data, and any relationships not fixed by Relationship Algorithm 1 were fixed in subsequent processes.

**Donor Imputation in CANCEIS:** CANCEIS can only process households of the same size together. For households size 1-5, relationships were imputed as part of the demographics module alongside relevant variables such as marital status and age. There were some rules about acceptable relationships and the connection to age and marital status (see Edit Rules, Section 4.5).

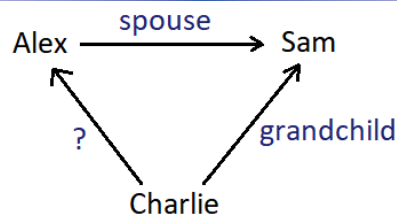
Since there were a smaller number of larger households, and as the dataset was split into ten geographical “Processing Units”, there were not enough households of size 6+ to attempt donor imputation. Persons 1-5 were imputed with 5-person households, and persons 6+ were imputed as individuals. Relationships could not be imputed this way.

**Relationship Algorithm 2** was a copy of Relationship Algorithm 1, designed to be applied after donor imputation. However, it did not change any records since the edit rules in donor imputation prevented any inconsistencies from being created by imputation. It was decided that Relationship Algorithm 2 was redundant and so it was removed from Edit and Imputation for 2011 live processing.

**Relationship Algorithm 3** was a deterministic algorithm which imputed relationships for persons 6+. It used triangulations, as well as relevant variables such as age and marital status, to best impute relationships.

For example (Figure 3), suppose Alex is the spouse of Sam, and Charlie is the grandchild of Sam. If Alex is at least 24 years older than Charlie, then Charlie would be imputed as the grandchild of Alex. Otherwise, Charlie would be imputed as unrelated to Alex.





**Figure 3: Example of a relationship triangulation:  
Deducing the relationship between Charlie and Alex**

#### 4.5 Edit Rules

An edit rule is a rule which determines an inconsistency or outlier. A full list of edit rules from 2011 can be found in Section 8.4.

#### Hard Edits

A **hard edit** is something which is impossible or so rare that most occurrences are errors. A hard edit specifies things which we will not allow in the dataset. For example:

- *A person under the age of 17 cannot drive to their place of work or study*
- *A person cannot have more than one spouse in the household.*

The first rule is straightforward. However the second rule is a remote possibility. However, while polygamy is legal in some countries, it is not legal in the UK and hence so rare that most instances in the dataset are in fact respondent errors.

Suppose there are only, say, 500 cases of a characteristic in all census responses throughout Scotland, and 495 of these cases are definitely errors (the remaining five being possible but not certain). It is not good enough quality to leave all 500 cases in the dataset, however it is impractical to manually check all of these cases to judge whether they are genuine, and such decisions could be biased (e.g. making judgements based on country of birth as to whether a person is in a polygamous

marriage). For these reasons, the hard edits were applied across the whole census dataset.

## Soft Edits

A **soft edit** is something which is very unlikely, which we wish to keep in the dataset but we do not wish to create disproportionately through imputation. For example:

- *A person is unlikely to be more than 65 years older than their child.*

It is possible that a person over the age of 65 could be a parent, for example they had biological children at a late age, or they adopted (note we expect a lot of adoptive grandparents to list their relationship as grandparent rather than parent). However, unless we explicitly tell the software that this is very unlikely, it does not understand the significance of age to parentage, and may impute a missing grandparent relationship as a parental one.

Note we can tailor soft edits so that they either don't create any new outliers through imputation, or only create an outlier when the donor record is also an outlier, which means that the distribution of outliers in the dataset is preserved.

## 4.6 Administrative Data

Administrative data is data which is collected primarily for administrative purposes, but which can be repurposed for statistical purposes, through data sharing agreements. For example, GP registration data contains names, addresses, and dates of birth of people registered at GP practices. The primary purpose is for the administration of patients and healthcare. Use of administrative data for statistical purposes is governed by [General Data Protection Regulation](#) (GDPR).

However, this represents a potential source of information on how many people live in an area, and what age they are. It will not capture everyone - for example, people who have moved in and out of the area but have not updated their details, or people

who have not registered with a GP. There may be an under or overestimation to this, for example young people with no health problems may not register locally until it is necessary or a person has forgot to de-register when moving abroad. But it generally can give a good indication of the usual resident demographic.

We did not use administrative data in 2011 to enhance Edit and Imputation, however administrative data sources such as the electoral register were used as part of quality assurance to compare distributions before and after imputation, such as age-sex by local authority. If the imputed census distributions were not similar to the administrative data distributions, the data and processes could be examined to look for potential introduction of bias.

#### 4.7 Voluntary Questions

The only voluntary question in 2011 was the question on religion. As a voluntary question, we expected that some respondents would choose not to answer, and it was not imputed on counted responses.

However for synthetic records added through Estimation and Adjustment (unit imputation mentioned in section 3), religion was imputed as part of the culture module, but “missing” was treated as a valid value.

Imputation of voluntary questions on synthetic records ensured that the distribution of responses in aggregate tables was preserved and not diminished by unit imputation.

For example, suppose that 10% of respondents indicated in a voluntary question that they identified as belonging to category X. Suppose we estimated that we had received responses from 95% of Scotland's population. When we added in synthetic records for the remaining 5%, if we were to leave the response to that question blank, it would appear in our main output tables that 9.5% (i.e. 10% of 95%) of Scotland's population identified as belonging to category X.

This is the difference between a count and an estimate: In this example we counted responses in category X from 9.5% of the true population, but we estimated, based on our estimate of the number of responses missed, that in fact 10% of the population would have indicated that they belonged to this category, had we received a response from everyone. This prevents casual data users from mistakenly inferring that only 9.5% of respondents identified as belonging to category X.

#### 4.8 Ad-hoc data adjustments

Although these were avoided where possible, it was sometimes necessary to make a deterministic data adjustment, called a Data File Amendment. This was initiated using a Request For Change which explained the problem and proposed solutions. This was then discussed and a solution agreed using a panel of subject matter experts.

These were not limited to Edit and Imputation, but were used from the point at which we received the coded data from our suppliers, to the various releases of output data following quality assurance work.

#### **An example of a Data File Amendment for Edit & Imputation:**

The long-term conditions question had a high non-response rate in 2011, which was believed to be caused by respondents not seeing the “No condition” option at the bottom of the question (Figure 4), and so skipping the question entirely. The non-response rate was so high that too many records required imputation, which affected the donor pool (see Section 3.3) and resulted in records failing to impute.

The data file amendment changed the missing response to “No condition” for respondents who otherwise indicated that they had good or very good health, and no disability.

20 Do you have any of the following conditions which have lasted, or are expected to last, at least 12 months?

◆ Tick all that apply.

Deafness or partial hearing loss

Blindness or partial sight loss

Learning disability (for example, Down's Syndrome)

Learning difficulty (for example, dyslexia)

Developmental disorder (for example, Autistic Spectrum Disorder or Asperger's Syndrome)

Physical disability

Mental health condition

Long-term illness, disease or condition

Other condition, please write in

or

No condition

Figure 4: Long-term Conditions question on paper form in 2011

#### 4.9 Quality Assurance

Some examples of the sorts of quality assurance checks used in 2011:

- Check changes made by Relationship Algorithms
- Counts of soft edit failures and use as donors
- Distribution of imputed values
- Frequency of hard edits
- Check imputation process, rates, etc. from CANCEIS diagnostics
- Counts of population sub-groups

Graphs and charts were produced using Excel to carry out this quality assurance.

#### 4.10 Audit and Metrics

The **non-response rate** for a variable is the proportion of submitted responses with missing or invalid values for that variable.

Variable non-response rates were published in the Scotland's Census 2011 General Report [6]. The lowest non-response rates were for age and sex (0.7% and 0.8%

non-response respectively) and the highest non-response rates were for last year worked and long-term conditions (16.8% and 15.2% respectively).

A full table of item non-response rates from the Scotland's Census 2011 General Report can be found in Annex 8.6.

There were also imputation flags available which indicated if a value had been imputed. These flags are very useful for internal quality assurance, as we can track the changes made to the dataset at each stage. The flags can also be useful for data users who want to work with data which has not been imputed. Imputation flags were not published with the main aggregate output tables, but were potentially available for record-level data extracts (such as the extract for the Scottish Longitudinal Study) depending on the level of disclosure control required. Imputation, although not the main part of the statistical disclosure control strategy, adds an extra level of uncertainty to small numbers in the output datasets and tables.

## **5. Proposed Methodology for 2022**

We are committed to using CANCEIS for donor imputation in 2022. The software is used, and thus tested, internationally.

There has been a major update since 2011 which makes CANCEIS more user-friendly. This is the use of spreadsheets to specify inputs, instead of collections of text documents.

Despite the improvement to the input interface, outputs are still mainly tab-delimited text files without column headers. In NRS we have developed “wrappers” to convert these files into usable datasets with headers and formatting, which are required for further processing steps. This will help us troubleshoot errors and monitor data quality much more easily.

The newer versions of CANCEIS are faster to run, and make use of threading, where a computer can process multiple records simultaneously. This makes processing much faster than in 2011.

## 5.1 Modularisation

Work on modularisation is not yet complete, and there are a number of decisions yet to be made, as outlined below. The modularisation will be based on the 2011 strategy with the following changes:

- New question on passports will be imputed in culture module
- New question on ex-service members will be imputed in labour market module
- New question on British Sign Language will be imputed in culture module or health module - more research required.
- We will be considering whether carer question would be better grouped with labour market questions, if the labour market variables make better predictors for the carer question.
- We will be considering whether economic activity question would be better grouped with labour market questions, if the labour market variables make better predictors for the economic activity question.
- Labour market module will be imputed as individuals rather than as households (increases data quality, improves processing time)
- We will be considering whether to impute health module as individuals rather than as households, if the health of other people in the household is less relevant than one's own health when imputing an individual's health questions.
- Partial codes (Postcodes, Standard Industrial Classification (SIC) codes for industry and Standard Occupation Classification (SOC) codes for occupation) will be imputed in CANCEIS as separate modules
- Pulling only the variables required for each module through CANCEIS, rather than pulling every variable in the census dataset through every module,

reduces processing time by about one-third, without affecting the methodology or the outcome of imputation.

- Choice of predictor variables, and weighting strategy for predictor variables, will be reviewed.

The following table is a working draft of the modularisation for 2022. There is still work to be done before this is finalised. Items highlighted in red are new to the questionnaire in 2022, or are existing questions which we are considering whether to impute in different modules.

Demographics	Culture	Health	Labour market
Age	Address 1 year ago	Carer	Qualifications
Sex	Country of birth	Disability	Ever worked
Marital status	Date arrived in UK	Health	Hours worked
Full-time student	Ethnicity	Long-term	Employee status
Term-time location	National identity	conditions	Supervisor
Relationships	Language questions	<b>BSL here</b>	Industry
Economic activity	<b>Passports</b>	<b>instead?</b>	Occupation
	<b>British Sign Language (BSL)</b>		Work/study address
			Method of travel
			<b>Ex-service</b>
			<b>Carer here instead?</b>
			<b>Economic activity here instead?</b>

**Table 5: Modularisation of non-voluntary person variables in 2022:  
Where variables will be imputed**

We considered allowing variables to be imputable after the main module in which they are imputed, but this is impractical as it can create inconsistencies with variables which have already been imputed. For example, although age is first imputed in the demographics module, it may be that there are contradictions with the age in the labour market module. However if age is changed in the labour market



module, this may introduce inconsistencies with previously imputed variables such as date of arrival in the UK.

We can however include variables which have not yet been imputed as predictors in a module, as was done in 2011 when country of birth was used as a predictor in the demographics module, but it was imputed in the culture module. Variables can also be used as predictors in subsequent modules, such as age being included in all four main person modules. We will be reviewing the choice and weighting of predictor variables in upcoming methodological research.

Response mode (paper/online) will continue to be used as a predictor variable.

There are significant demographic differences between people who respond online and people who respond by paper, and since paper responses tend to contain more error, and we expect the majority of responses to be online, there is a risk that the demographic groups responding by paper will be matched with donors from online responses, who are not part of the same demographic subset of the population [7]. The inclusion of response mode as a predictor variable reduces this risk.

## 5.2 Processing Units

We have decided not to split the data into processing units for Edit and Imputation of Scotland's Census 2022. The entire Scotland dataset will be processed in the same batch.

If we wish to geographically limit the distance between a failed record and its potential donors, we can do this by changing the parameters in CANCEIS to limit how far the software will search for donors.

We have more powerful computers now than in 2011, and the CANCEIS software is faster to run. Although splitting the dataset into 10 processing units and running these simultaneously would speed up the process, the actual run-time of these processes is insignificant, with the absolute maximum of any one CANCEIS module taking about 2 hours to run on all Scotland. By comparison, quality assurance

checks after the completion of each process will take much longer, and splitting the dataset for quality assurance checks does not save any time.

The main advantage of processing all Scotland at once is that we can impute larger households together (see Section 4.2), which improves data quality. There will still be a natural limit to the size of household we can impute in this way, so inevitably there will be a cut-off point at which we will have to impute household members as individuals, but this cut-off point will be much larger than in 2011, so there will be fewer household members who need individual imputation. As we explain in Section 5.4, this allows us to improve our imputation of relationships as we no longer need Relationship Algorithm 3.

### 5.3 Partial codes

We now have a better understanding of how CANCEIS works since we used it in 2011, and so we have developed a way of using CANCEIS to impute full postcodes by donor imputation, using partial postcodes as predictors, as well as other relevant questions. For example, industry can be useful to pinpoint similar places of work.

We can also impute full codes for industry and occupation using similar methodology. This was not considered in 2011: industry and occupation codes were assigned in full or considered “invalid”. For example, if a person said they were a teacher, but it could not be determined at what level they taught (primary, secondary, further, higher) because there was no employer information, then their occupation was coded as “invalid” and the information was lost. Since there was no employer information, it is unlikely that their imputed occupation would be a type of teacher.

While we cannot say exactly how many partial codes could have been assigned, approximately 53,000 records (1.4% of responses where something was entered) contained an invalid value for occupation. Some of these could have been assigned partial codes, while others would be invalid for other reasons.

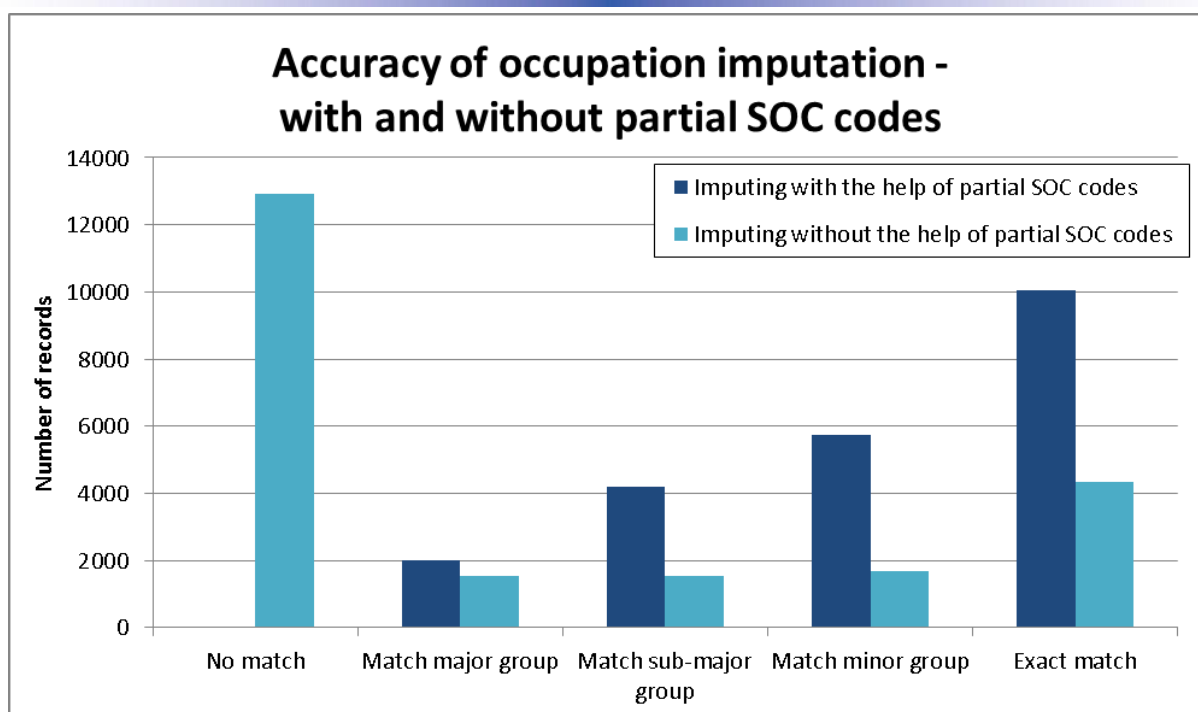
Like postcodes, industry (SIC) and occupation (SOC) codes are hierarchical, so we may be able to glean some higher level of information from a response: for example, if a person says that they are a teacher but they have not specified at what level, we can look at their place of work, their level of qualifications, etc. to predict whether they are a primary, secondary or higher education teacher.

Figure 5 illustrates the potential increase in data quality when partial SOC codes are used to help impute full SOC codes. This is taken from a test in which approximately 6% of records from a test dataset (n=350,000) were selected to change the full SOC code given by the respondent into a partial code. The original, full code was compared to the code assigned by imputation. The chart shows whether the original and imputed codes matched exactly, not at all, or matched on one of the hierarchical partial SOC groups: major group (lowest level of information), sub-major group, or minor group (highest level of information).

The two methods being compared were:

- Partial codes not used as predictors - 2011 method (light blue bars)
- Partial codes used as predictors - proposed method (dark blue bars)

There was a significant improvement in the accuracy of imputation when using partial codes as predictors (dark blue bars), compared to the 2011 method of imputing without that information (light blue bars). Significantly, there were more than twice the number of exact matches when using partials information, and there were almost no cases where the imputed occupation did not match the original occupation at all.



**Figure 5: Imputation using partial codes (dark blue bars) significantly improves accuracy and data quality, more closely matching the “true” occupation value, compared to imputation without partials as in 2011 (light blue bars).**

**This is a test dataset which is not indicative of the number of responses which could have been assigned a partial code in 2011: this number is unknown.**

There was also an issue of consistency between industry and occupation as a result of imputation in 2011. For 2022 we plan to use conditional parameters in CANCEIS which increase the predictive weight of industry or occupation when either variable needs to be imputed.

For example if occupation is missing but industry is given by the respondent, then the predictive weight of the industry variable is increased so that CANCEIS prioritises potential donors working in the same industry. This significantly decreases the likelihood of imputing a unique industry-occupation combination. This is a conditional parameter, so if industry and occupation do not need to be imputed, then their weights are not increased when imputing other variables in the labour market module. It is not vital, for example, that a donor record is from the same industry or occupation when imputing method of travel to work.

#### 5.4 Relationships

The relationships question in the questionnaire [8] has been improved for 2022 to help respondents fill it in:

- In-question guidance added for in-laws (other relation)
- Half-siblings (sharing one parent) are now classed with siblings (two parents in common) rather than step-siblings (only step-parents in common)
- Name field on paper questionnaires was not captured in 2011. For 2022 it will be captured to help match individual questions (age, sex, marital status etc.) with relationships
- Online questionnaire is easier to fill in: it uses the format <person name> is [choose relationship] to <person name>, and relationships already filled in other direction are displayed for checking (if Alice is parent of Bob, then for Bob's questions it shows that Bob is child of Alice).

Together with responses being largely online, this should result in an increase in incoming quality of relationships, which will increase the quality of imputed relationships.

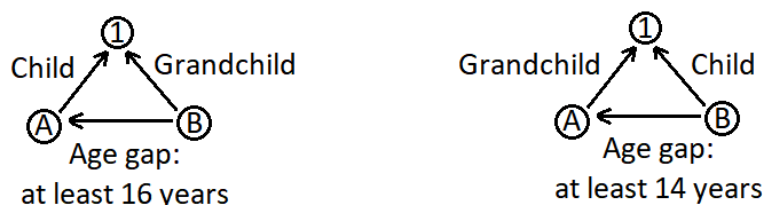
Relationship Algorithm 1 will be retained for 2022. As in 2011, Relationship Algorithm 2 will not be used as it will not have any effect on the data.

Relationship Algorithm 3 was a very large, undocumented piece of code in 2011, bringing issues of a lack of transparency. Furthermore a quick analysis revealed that relationships were not treated symmetrically as would be expected, so the imputed relationship depended on what order the three people appeared in the dataset. As a result, there were doubts about the transparency and accuracy of this algorithm.

For example, Table 6 shows a condition involving a child and grandchild of another person (person 1). Depending on whether the child or the grandchild is listed first, the age gap requirement to impute a parent/child type relationship is either 16 years or 14 years.

<b>Relationship of person A to person 1</b>	Child	Grandchild
<b>Relationship of person B to person 1</b>	Grandchild	Child
<b>Condition</b>	$A - B \geq 16$	$B - A \geq 14$

**Table 6: Example of a condition which was not applied symmetrically in RA3**



**Figure 6: Example of asymmetric imputation conditions**

Relationship Algorithm 3 was only applied to household persons 6+ in 2011, whereas the demographics donor imputation module imputed relationships for persons 1-5, meaning that two different methods were being used to impute relationships for different individuals.

In addition, the decision to process all Scotland together through Edit and Imputation in 2022, instead of splitting the dataset into processing units, allows us to impute larger households (see Section 5.2). This increases the size of households which can be imputed in the main donor method, and reduces the number of household members requiring individual imputation.

As part of this development, there were extra edit rules developed regarding household relationships (see Section 5.5 for a current proposed list of new relationship rules). This reduced the number of implausible relationships imputed using donor imputation. When tested against Relationship Algorithm 3, the donor method produced fewer implausible relationships (see Table 7). Additionally the methodology is much easier to explain and understand, as well as being much easier to review, develop and test.

Method	Frequency
RA3	63
CANCEIS (1 <sup>st</sup> iteration – rules as 2011)	78
CANCEIS (3 <sup>rd</sup> iteration – extra rules as listed in Section 5.5)	89

**Table 7: Number of households with plausible imputation  
(out of 100 sampled for each method)**

The actual maximum size of household which can be imputed in this way will depend on the number of households of each size in 2022, so the threshold cannot be decided until we receive the data. In our documentation, we use an example threshold of households containing 10 usual residents, as this works well with the 2011 dataset in testing, since only approximately 140 households contained more than 10 usual residents. However the threshold could be different for the 2022 dataset.

As in 2011, we will impute the first 10 (or so) people from larger households together with individuals from households of size 10. Then persons 11+ will be imputed as individuals using donor imputation in CANCEIS for all other demographics variables (age, sex, student status, etc.), but the relationships cannot be imputed this way. We can use CANCEIS to detect and flag missing, invalid and inconsistent relationships for these remaining relationships, but decisions on how to resolve these relationships will have to be done by manual inference using the other household relationships, and supporting information such as age and marital status. The increase in the threshold from size 5 to approximately 10 will reduce the number of manual interventions required, making this feasible without the need for an automated process like Relationship Algorithm 3.

If the 2011 data had been processed using the new method, approximately 34,000 individuals would have been imputed as part of their household, instead of being imputed as individuals. The remaining individuals (only persons 11+) from the approximately 140 remaining households would have to be imputed deterministically. This is a small enough number that the process doesn't need to be automated.

## 5.5 Edit rules

Changes to the edit rules will be driven by changes to the questions, changes in society (e.g. legalisation of same-sex marriage), and improvements to the quality of data.

We work closely with ONS and Northern Ireland Statistics and Research Agency (NISRA) to ensure harmonisation of outputs, and the use of hard edit rules is the main place in Edit and Imputation where harmonisation matters, as these will determine what is not allowed in the dataset. It is important that data users do not draw their own conclusions as to why a characteristic appears in, say, English and Welsh data and not in Scottish data. NRS will publish a list of their edit rules for 2022 when they are finalised.

The main changes for edit rules in Scotland are at present:

- “Unless country of birth is outside the UK” clause is being dropped (e.g. “A person under 16 cannot be, or have been, married, unless country of birth is outside UK”)
- Marriage & Civil Partnership (Scotland) Act 2014: Same-sex marriage now possible
- Civil Partnership (Scotland) Bill: Opposite-sex civil partnerships to become possible
- Some questions such as marital status, address 1 year ago and carer have new age routing and so come under filter rules [9] rather than edit rules. For example, edit rules about marital status for under-16s are no longer relevant as this question will not be asked of under-16s.

As part of the retirement of Relationship Algorithm 3 in 2022, we have developed extra relationship rules. As with all the edit rules involving households, these rules are for people living together as a household at the same address. Some respondents would confuse the columns in the relationship matrix (Figure 2), resulting in their children being misreported as their partners, and so on. Imputation



of relationships is highly complex, and the signal-to-noise ratio of erroneously reported relationships and unusual relationships within a household is too high to be able to retain many highly unusual relationships (such as a person living with both their spouse and their partner).

This list of rules is still work-in-progress, still under review, but are currently as follows:

- If two people share a parent, then they are (half) siblings
- Two people can only share a step-parent/parent if they are (half) siblings or step-siblings
- If two people share a sibling or step-sibling within a household, then they are siblings or step-siblings
- (Soft Edit) If two people share a child or step-child, then they are unlikely<sup>5</sup> to be anything other than partners/spouses/civil partners, unless marital or civil partnership status is divorced, dissolved or separated
- Two people can only share a child, step-child or grandchild if they are partners/spouses/civil partners, or ticked the “other relation” or “unrelated” categories.
- If two people are spouses/civil partners then the child of one must be the child or step-child of the other
- If two people are partners then the child of one must be the child or step-child of the other, or unrelated to the other. For example the child of one person cannot be their partner’s sibling.
- (Soft Edit) If two people share a grandchild, then they are unlikely<sup>6</sup> to be anything other than partners/spouses/civil partners
- Two people can only share a grandparent if they are siblings, step-siblings or cousins

---

<sup>5</sup> In 2011 data, approx. 1.1% of cases where two people shared a step-child, were anything other than partners/spouses/civil partners.

<sup>6</sup> In 2011 data, approx.. 4.3% of cases where two people shared a grandchild, were anything other than partners/spouses/civil partners

- A person cannot have more than one spouse/civil partner/partner in the household
- A person cannot have more than two parents
- Three-generation triangulation (parent/child/grandchild, where one of these relationships is missing)
- (Soft Edit) There is unlikely<sup>7</sup> to be more than a 20-year age gap between siblings

We are also considering greater use of soft edits (outliers we do not wish to disproportionately propagate), such as these, where the proposed age thresholds are based on the 1<sup>st</sup>/99<sup>th</sup> percentiles of records in the 2011 dataset:

- A person is unlikely to be retired until the age of 53
- A person is unlikely to be widowed below the age of 42
- Full-time students are unlikely to be more than 45 years of age
- Students (full- and part-time) are unlikely to be more than 55 years of age
- A non-staff resident of a (age-specific communal establishment, e.g. retirement home, boarding school) is unlikely to be above/below the age of X. (Still in development)

## 5.6 Administrative Data

As in 2011 we will continue to use administrative data sources as part of data quality assurance, to compare aggregated administrative data to census data. For example, we may compare age groups in each local authority using sources such as mid-year estimates. This will be developed as part of the statistical quality assurance strand of work. We expect there to be some differences between the census data and administrative sources, but significant differences will be investigated to see if there is an issue with the census data or the imputation process.

---

<sup>7</sup> In 2011 data, in approximately 0.9% of cases where two people were siblings, was the age gap more than 20 years.

Additionally, we aim to use administrative data to enhance Edit and Imputation in 2022. The admin data team will link the census dataset to the administrative dataset and compare dates of birth. Where the census date of birth is missing or different from the administrative date of birth, we will receive a dataset containing the census identifiers with the age from the administrative dataset. This information attached to each census record aims to provide additional information to aid the imputation process [10], as explained below.

Where the census age is missing or different from the admin age, we will not directly copy the value from the admin dataset into the census responses, but it will help inform selection of donor information where the census age is missing or inconsistent, using a method developed by the Office for National Statistics [11].

Age is a very important variable for predicting other census responses, as well as being one of our key output variables. Using administrative data increases the accuracy of imputed age.

Table 8 and Table 9 is an example of how this works. Administrative age is provided for all matched records where it is different from the census age, or where the census age is missing. For all other records, the value for administrative age has been copied from the census age<sup>8</sup>. Record 2 is missing the census age, but administrative data suggests that the age is 25. Of this small dataset, record 1 has the closest administrative age, and along with other variables not shown, this is the most similar record to record 2. Record 1 is selected as a donor and the age, 23 years, is imputed.

---

<sup>8</sup> We make no distinction between census records which matched with administrative data records, and those which did not match. This is because the absence of a match to administrative data does not indicate that the given census age is inaccurate. Instead, it may, for example, be due to the individual not being included in the administrative dataset, or the individual giving different forms of their name, or having moved house recently.

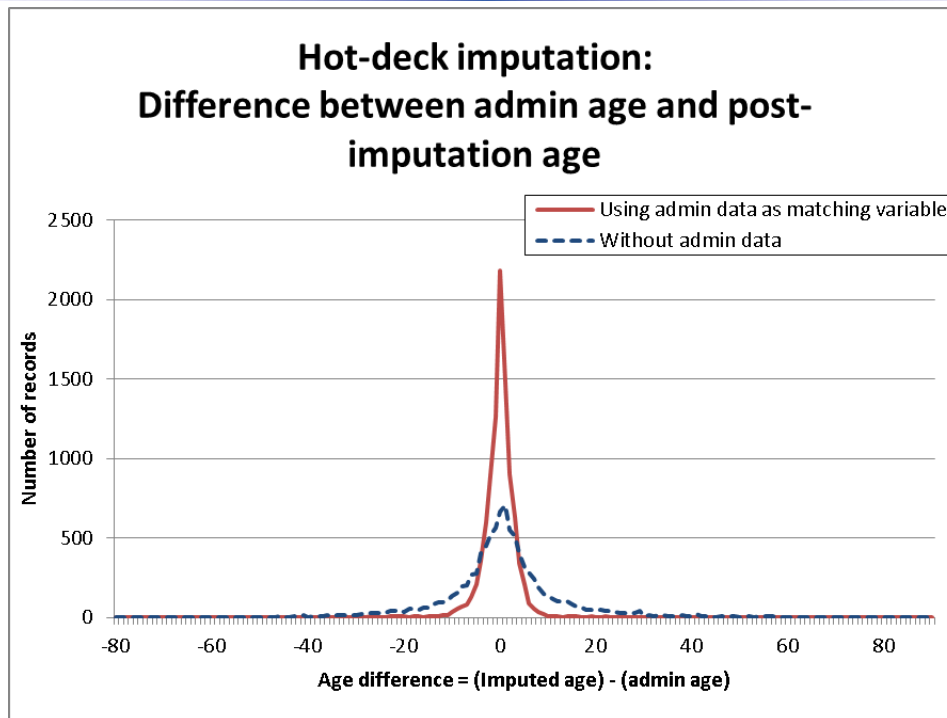
Record ID	Census Age	Administrative Age	Other variables...
1	23	23	...
2	missing	25	...
3	56	56	...
4	8	8	...
5	74	74	...

**Table 8: Example of donor imputation with administrative data  
Before imputation (This is fake data)**

Record ID	Census Age	Administrative Age	Other variables...
1	23	23	...
2	23	25	...
3	56	56	...
4	8	8	...
5	74	74	...

**Table 9: Example of donor imputation with administrative data  
After imputation (This is fake data)**

Often the imputed age be the same as the administrative data age - this is good, because we want the imputed age to be accurate, and tests on the administrative age suggested that it is more accurate than the census age. Sometimes the imputed age will be different to the admin age, but in this case it will usually be close, certainly much closer than without the use of administrative data. Figure 7 shows the results of testing this imputation method using 2011 data. The use of administrative data was shown to produce much more accurately imputed age.



**Figure 7: Comparison of imputed age to admin age:  
With and without use of administrative data**

## 5.7 Voluntary Questions

There are two new voluntary questions proposed for 2022: sexual orientation and trans status or history. As with the voluntary question on religion in 2011, these questions will only be imputed on synthetic records added in estimation and adjustment, and “missing” will be included as a valid option for imputation. This will increase the observed distributions proportionally as synthetic records are added to the dataset, as explained in Section 4.7.

The modularisation of these questions has not been agreed yet, however it is expected that the new voluntary questions will be imputed in a new, separate module. It has not been decided whether we want to include age, sex etc. as “predictors” to preserve key distributions, since data may not be missing at random.

## 5.8 Ad-hoc data adjustments

We will need to have a process for live running to make ad-hoc deterministic changes to the data.

Particularly relevant to Edit and Imputation will be any questions with high levels of non-response which result in a large proportion of records requiring imputation.

It is important for donor imputation that there are plenty of records which pass the edit and do not need to be imputed: this is the pool of potential donors. The more potential donors there are per failed record, the more likely it is that the software can find a very similar record to use as a donor, which improves plausibility, and hence quality, of imputation.

Although there is no set optimal ratio of failed records to potential donors which guarantees successful imputation, we can monitor factors such as the number of records failing imputation (where a suitable donor could not be found) and the number of times donors are reused to help us decide whether there are enough donors. We can also limit the number of times each record can be used as a donor to prevent clusters of similar looking records.

Considering the example from 2011, the high non-response to the long-term health conditions question: evaluation of rehearsal data is ongoing, but so far there are indications that the “online first” approach leading to the majority of returns being electronic submissions, will help mitigate this risk, as the validation messages reminding respondents to answer the question appear to be reducing non-response.

## 5.9 Quality Assurance

We are still developing our quality assurance strategy [12] for 2022 and in particular we are currently developing a data visualisation tool which will help us produce graphs and charts quickly and easily, making the quality assurance process much more efficient. We are using the open source software R, familiar to many

statisticians, specifically the Shiny package [13] which allows us to build a browser-based interactive dashboard. The work is still in development and will have to be tested and signed off prior to live census.

The dashboard is not part of our public-facing outputs. The purpose of these graphs and charts is not for publication, but to check that processes have been performed correctly and that we have not introduced bias into the data. The majority of quality assurance will be completed internally. However the data may be reviewed by a panel of subject-matter experts before key processes are signed off. The sign-off process has not been agreed at this stage.

We aim to produce graphs to show:

- Single distributions and cross-distributions at important geographic levels (Local authorities etc.) e.g. single year of age by sex per local authority.
- Comparisons before and after imputation
- Census data against comparator sources (e.g. mid-year population estimates)

There will be particular potential issues which we will look for when processing 2022 data, based on what we have learned from 2011 data. In many cases we expect these issues to have been resolved through re-design of the questions or guidance messages, refinement of processes, or by virtue of most responses expected to be online.

For example, in 2011 there was a spike in the age distribution at the 15/16 years mark. This was due to the software “nudging” the individual’s age just over the boundary to make it consistent with other responses (marital status, for example). This was considered by the software to be a smaller change than changing the other value(s).

CANCEIS has functionality to set a “threshold” age in the distance metric, so that moving the age across that threshold counts as a large change, which should sort out this problem. Additionally, we expect most responses to be online, where

calculated age is displayed to the respondent when they enter date of birth, and questions which should not be answered by under-16s are automatically routed around.

However we will still want to monitor this for 2022 live processing to ensure that the issue does not reappear despite these changes.

## 5.10 Audit and Metrics

### **Imputation and non-response rates**

The non-response rate was defined in section 4.10. The **imputation rate** for a variable is a slightly different concept. It is the proportion of submitted returns where that variable has been imputed due to missing or invalid values, or inconsistencies.

Since the non-response rate does not include inconsistencies, it will be lower than the imputation rate. Both are dependent on the quality of the input data: the non-response rate is a measure of how many people did not answer, or gave an invalid response to, a question. The imputation rate includes all these for mandatory questions, as well as inconsistent responses which may be due to respondent error or misunderstanding (e.g. reporting parent-child relationship the wrong way round), or may be due to process errors (e.g. the numbers in a date of birth field on a paper return are incorrectly interpreted by the automatic software, such as mistaking a 0 for a 6).

We will again publish non-response rates for 2022 comparable to 2011 published rates. For process quality assurance (not necessarily for publication), we will provide the following summary figures:

- Number of records flagged by each edit rule
- Number of records imputed for each variable
- Number of records with missing or invalid values for each variable



Since the non-response rate and imputation rate are two similar, but slightly different concepts, we may choose not to publish both, to avoid confusion. If we decide to publish the imputation rate, then it should be made clear to external users whether deterministic changes are included in this count.

## Imputation flags

We can produce flags for every record, in every variable, that say whether the value was imputed or provided by the respondent. We can break these down by process if required. These flags will continue to be a very useful tool for process and data quality assurance, and will be a record of what was done to the data.

Imputation flags will also be useful in some circumstances, for researchers who wish to work on record-level extracts of the census dataset without imputed values.

Imputation is a process which has positive implications for Statistical Disclosure Control, as data users will be unaware of whether an individual value is observed from a return, or imputed. Therefore there will be restrictions on access to these flags:

- Flags will not be supplied with aggregate tables (such as the main outputs on our website).
- Scottish Longitudinal Study record-level extract: This will contain record-level imputation flags.
- Safeguarded extracts (such as origin-destination data) will have statistical disclosure control applied and therefore imputation flags will not be supplied.
- Bespoke extracts will not typically include imputation flags, but potentially could. Privacy implications would be assessed at the application stage.

## 6. Conclusions

The Edit and Imputation methodology for Scotland's Census 2022 builds on the robust statistical process from Scotland's Census 2011, with improvements which

increase the quality of census outputs and improve processing time, helping to deliver on our goal of first outputs within a year of Census Day.

In particular, improvements to computing power and software allow us to make better use of the CANCEIS donor imputation software, using the power of larger datasets to enable greater use of donor imputation for larger households, in order to improve accuracy and process transparency, and ensure that a much larger proportion of the population had relationships imputed using the same method (donor imputation).

We are also making better use of hierarchical codes such as postcodes and SIC and SOC codes, to enable us to use partial information provided by respondents, where a full code cannot be identified, in order to more accurately assign a value through donor imputation. Linked to this, we also have improved imputation of industry and occupation codes to prevent the creation of implausible industry-occupation combinations.

Modularisation is under review, to ensure that we are imputing variables with the best predictors, particularly with the questions on unpaid care and economic activity. We will now impute labour market questions as individuals rather than as households, as this will improve accuracy, with a considerable improvement to processing time being an added bonus. This is being considered for the health module as well, but further analysis is needed before a decision can be made.

We are making use of administrative data for the first time to help improve the accuracy of age imputation. Where the census age is missing or different from the admin age, admin age will be used, along with other predictors, to find records of a similar age. This age will be copied from the census response of the donor record into the census response of the failed record.

We are currently developing a data visualisation tool which will allow us to review the imputed dataset and compare it with the pre-imputation responses, as well as alternative sources of data, as part of process and data quality assurance. The

expected increase in online responses will provide higher quality data being fed into the imputation process, which in turn increases the quality of imputation.

The edit rules are being updated to reflect the new and changed questions, and we are working closely with the other UK statistics offices ONS and NISRA in order to ensure harmonisation of UK outputs.

## 7. References

- [1] National Records of Scotland, "Estimation and Adjustment Methodology," 2020. [Online]. Available: [https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20\(pdf\).pdf](https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf).
- [2] I. Fellegi and D. Holt, "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, vol. 71, no. 353, pp. 17-35, March 1976.
- [3] M. Bankier, "Experience with the new imputation methodology used in the 1996 Canadian census with extensions for future censuses," in *UNECE Conference of European Statisticians: work session on statistical data editing*, Rome, 1999.
- [4] M. Bankier, "2001 Canadian Census minimum change donor imputation methodology," in *UNECE Conference of European Statisticians: work session on statistical data editing*, Cardiff, 2000.
- [5] Office for National Statistics, "2011 UK Census: An overview of the edit and imputation process," in *UNECE Conference of European Statisticians: Work Session on Statistical Data Editing*, Ljubljana, 2011.
- [6] National Records of Scotland, "Scotland's Census 2011 General Report," October 2015. [Online]. Available: [https://www.scotlandscensus.gov.uk/documents/censusresults/Scotland's\\_Census\\_2011\\_General\\_Report.pdf](https://www.scotlandscensus.gov.uk/documents/censusresults/Scotland's_Census_2011_General_Report.pdf). [Accessed 7 July 2020].
- [7] S. Rogers, L. Dyer and B. Foley, "Towards the 2021 UK Census imputation strategy: Response mode as a matching variable in a donor-based approach?," in *Proceedings of Statistics Canada Symposium 2014*, 2014.
- [8] National Records of Scotland, "Question Set for Scotland's Census 2022," 2020. [Online]. Available: <https://www.scotlandscensus.gov.uk/2022-question-set>.
- [9] National Records of Scotland, "Data Cleansing," 2020. [Online]. Available: <https://www.scotlandscensus.gov.uk/data-cleansing>.
- [1] National Records of Scotland, "Missing and Different Dates of Birth," 2020. 0] [Online]. Available: <https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0>.
- [1 F. Leather, K. Sharp and S. Rogers, "Towards an integrated census- 1] administrative data approach to item-level imputation for the 2021 UK Census," Neuchâtel, 2020.

- [1] National Records of Scotland, "Scotland's Census 2022 Statistical Quality Assurance Strategy," 2019. [Online]. Available: <https://www.scotlandscensus.gov.uk/documents/Statistical%20Quality%20Assurance%20Strategy.pdf>.
- [1] R Studio, "Shiny from R Studio," [Online]. Available: <https://shiny.rstudio.com/>.  
3]
- [1] National Records of Scotland, "Developing a Hard-to-Count Index," 2020.  
4] [Online]. Available: <https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0>.

## 8. Annex of supporting information

### 8.1 Definitions

**NRS:** National Records of Scotland

**NISRA:** Northern Ireland Statistics Research Agency

**ONS:** Office for National Statistics

**Edit:** The detection of missing, invalid or inconsistent responses.

**Imputation:** The correction of missing, invalid or inconsistent responses.

**Non-response rate:** For a variable is the proportion of submitted returns with missing or invalid values for that variable.

**Imputation rate:** For a variable is the proportion of submitted returns where that variable has been imputed due to missing or invalid values, or inconsistencies. It should be made clear to external users whether deterministic changes are included in this count.

**Item-level imputation:** Imputation of variables (items) within a record.

**Unit-level imputation:** Insertion of new records into a dataset (This comes under Estimation and Adjustment, and is not part of the work which NRS call Edit and Imputation.)

**Missing response:** Where a respondent has not answered a question, and a response was required.

**Invalid response:** Where a response was provided, but it is not an acceptable value.

**Inconsistent response:** A response to a question which contradicts other information given.

**Hard Edit:** A rule which defines something which is impossible or so rare that most occurrences are errors. A hard edit specifies things which we will not allow in the dataset.

**Soft Edit:** A rule which defines something which can be considered to be an outlier. A soft edit specifies things which we do not want to disproportionately propagate throughout the dataset as a result of imputation.

**Failed record:** A record which contains missing, invalid, or inconsistent responses.

**Donor record:** A record which 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, or resolve inconsistencies.

**Donor imputation:** Copying values from another “donor” record into the failed record.

**Hot-deck imputation:** Donor records come from the same dataset as the failed record.

**Fellegi-Holt imputation methodology:** imputation which is based on a series of self-contained edit rules and no defined imputation rules, which relies on the principle of minimal change to the dataset [2].

**Nearest Neighbour imputation methodology:** Donors are selected based on similarity to the failed record. There is also an assumption that geographically close records are demographically similar [3].

**CANCEIS:** Canadian Census Edit and Imputation System: Software designed by Statistics Canada, the primary function of which is to apply nearest-neighbour hot-deck donor imputation [4].

**SIC:** Standard Industrial Classification. Codes assigned to industry categories.

**SOC:** Standard Occupational Classification. Codes assigned to occupation categories.

## 8.2 “Similarity” in CANCEIS and the use of predictors

There are a range of distance functions available in CANCEIS, such as the discrete metric (two values have a distance 0 if they are the same and 1 if they are different), geographic distance (scaled to values between 0 and 1), use of bigrams for character variables, and so on.

The similarity between two records for individuals  $A$  and  $B$  is defined as

$$S_{A,B} = \sum_{i=1}^n w_i d_i$$

where  $n$  is the number of predictor variables,  $w_i$  is the weight for predictor variable  $i$ , and  $d_i$  is the distance between the values for variable  $i$  in the two records. The similarity between two household records  $X$  and  $Y$  is defined as the sum of the above for each individual in the household:

$$S_{X,Y} = \sum_{i,j} w_{i,j} d_{i,j}$$

for all variables  $i$  and household members  $j$ . These similarity scores are used when assessing potential donor individuals or donor households: potential donors with the lowest similarity scores are added to a shortlist, from which all possible imputation actions (choice of variables to impute) are considered.

When assessing a potential imputation action  $A$  between a failed record  $F$  and a potential donor  $P$ , the software calculates the similarity between the failed record and the record as it would look after imputation:  $S_{F,A}$ , and the similarity between the potential donor and the record as it would look after imputation:  $S_{P,A}$ . The first score is a measure of minimal change to the failed record, and the second score is a measure of plausibility of the imputed record.

There is a software parameter  $\alpha \in (0.5, 1]$  which can adjust the balance between the minimal change and plausibility scores for the overall imputation action quality score:

$$D_{F,P,A} = \alpha S_{F,A} + (1 - \alpha) S_{P,A}$$

Similarly to potential donors, potential imputation actions are shortlisted based on their  $D_{F,P,A}$  score. However when all potential donors have been considered and the shortlist of potential imputation actions is complete, one of these potential imputation actions is selected with probability inversely proportional to the imputation action quality score. Thus better imputation actions are more likely to be chosen to impute the failed record.

### 8.3 Relationship Algorithm 1: Common errors to fix

**ReIEdit1:** *Parent-child relationship has been reported wrong way round*

**ReIEdit2:** *Stepparent-stepchild relationship has been reported wrong way round*

**RelEdit3:** *Grandparent-grandchild relationship has been reported wrong way round*

**RelEdit4:** *Missing grandparent-grandchild relationship in three-generation group*

**RelEdit5a:** *Lone-parent family: missing sibling relationships*

**RelEdit5b:** *Two-parent family: missing sibling relationships*

**RelEdit5c:** *Stepparent family: missing sibling relationships*

**RelEdit6:** *Partner and (step)parent/sibling of a reference person should be related*

**RelEdit7:** *Two people who share a sibling or half-sibling must also be (step)siblings*

**RelEdit8:** *Two-parent families where two parent/child relationships are misreported as siblings*

## 8.4 Edit Rules in 2011

### 8.4.1 Household

- If the householders own their residence (outright or with a mortgage) then no response is required to the landlord question
- A household cannot live rent free if the landlord is the council, a housing association, or registered social landlord
- If there are no usual residents then number of cars should be zero

### 8.4.2 Demographics

- A person under 16 cannot be, or have been, in a civil partnership
- A person under 16 cannot be, or have been, married, unless country of birth is outside the UK
- A person under 16 cannot be, or have been, married



- A person aged between 5 and 15 must be a student in full-time education unless limited a lot by a health problem/disability
- A person cannot have more than one spouse/civil partner
- A person cannot have a spouse/civil partner and a partner
- Two people with at least one parent in common cannot be married/civil partners/partners with each other
- If two people are married then both cannot be of the same sex
- If two people are civil partners then both must be the same sex
- A person aged less than 16 cannot be a same sex civil partner
- A person aged less than 12 cannot be a parent
- A parent cannot be less than 12 years older than their child
- A person aged less than 16 cannot be a spouse unless country of birth is outside the UK
- A person aged less than 12 cannot be a partner/stepparent unless country of birth is outside the UK
- A person with a parent aged less than 28 must not have a marital status of married, separated, divorced or widowed unless country of birth is outside the UK
- A person with a parent aged less than 28 must not have a marital status of civil partner, separated civil partner, dissolved civil partner or widowed civil partner
- A person with a spouse in the household cannot have a marital status other than married or separated
- A person with a civil partner in the household cannot have a marital status other than in a civil partnership or separated but legally still in a civil partnership
- At least one usual resident must be 12 years old or above
- Person 1 cannot have more than two parents/stepparents

- A man is unlikely to be more than 65 years older than his child
- A person aged less than 24 cannot be a grandparent
- A grandparent cannot be less than 24 years older than their grandchild
- A person aged less than 10 cannot be a spouse
- A woman cannot be more than 66 years older than her child
- A person who is not working cannot have position = staff
- A response is not required for questions on relationship to self or subsequent household members

#### 8.4.3 Culture

- A person cannot arrive to live in the UK before their date of birth
- Person 1 cannot have response "same as person 1"
- A person who has ticked "speak" for English in Q16 cannot have "not at all" for how well they speak English
- A person who has ticked "no, English only" in Q18 must have ticked at least one of the English options in Q16
- Address 1 year ago variables consistency

#### 8.4.4 Health

- A person aged under 5 cannot be a carer

#### 8.4.5 Labour Market

- A person's year last worked cannot be before their date of birth
- A person who is not working cannot have a communal establishment position of staff
- A person under the age of 17 cannot usually travel to work/study by driving a car/van

- A person who is a full-time student cannot have work\_study\_address = not currently working or studying
- Workplace indicator and workplace postcode (also for study place)
- If working from home then a response is not required for the transport question

## 8.5 Processing Units in 2011

Council areas were grouped contiguously into processing units, each containing approximately 500,000 individuals, as follows:

- A. South Lanarkshire, East Lothian, Scottish Borders
- B. North, East and South Ayrshire, Dumfries & Galloway
- C. City of Edinburgh, Midlothian
- D. North Lanarkshire, West Lothian
- E. Fife, Clackmannanshire, Falkirk
- F. City of Glasgow
- G. Inverclyde, Renfrewshire, East and West Dunbartonshire, East Renfrewshire
- H. Angus, Perth & Kinross, Stirling, Dundee City
- I. Aberdeen City, Aberdeenshire, Shetland
- J. Argyll & Bute, Moray, Highland, Na h-Eileanan Siar, Orkney

## 8.6 Item non-response rates

Non-response rates for the rehearsal contain some caveats.

Rehearsal data only includes submitted paper and online forms. Unsubmitted online forms are not included due to data quality issues.

In some cases non-response rates are approximations, due to issues with quality of rehearsal data from paper responses.

Rehearsal data is not reflective of Scotland's population as a whole, and hence caution must be exercised when comparing rehearsal figures to 2011 figures. Similarly the demographics for online responses were different to paper responses. For example, a larger proportion of respondents of retirement age would result in a higher non-response rate for the student question, as respondents skip the question believing it to be irrelevant to them.

This table shows non-response rates for the rehearsal and for Scotland's Census 2011 (from the 2011 General Report) [6]. The lower overall non-response rates for each question in the rehearsal compared to 2011 is likely a result of two factors:

- 1) The rehearsal was a voluntary questionnaire - potentially people who did respond were more motivated to answer the questionnaire.
- 2) The majority of responses for the rehearsal were online, where item response rate was very high, perhaps due to validation messages reminding respondents to select a response.

Question	Rehearsal Online (Sumbitted)	Rehearsal Paper	Rehearsal total	2011 total
Accommodation type	0.0	2.4	0.3	1.7
Self-contained accommodation	0.0	2.6	0.3	1.3
Number of (bed)rooms	0.0	9.3	1.6	2.2
Type of central heating	0.0	1.8	0.2	1.6
Tenure	0.0	5.7	0.7	1.5
Landlord	0.0	5.4	0.7	1.8
Number of cars and vans	0.0	5.6	0.7	1.2
Relationship to person one	0.0	3.4	0.4	3.5
Sex	0.0	3.5	0.4	0.8
Age	0.0	13.5	1.7	0.7
Marital status	0.1	3.4	0.6	2.3
Student	0.0	16.7	2.1	5.5
Term-time	0.0	16.4	2.1	2.2
Country of Birth	0.0	3.2	0.4	2.0
Arrival in UK (year**)	0.4	3.6	0.8	5.1
Carer	0.0	6.1	0.8	2.9
Address 1 year ago (indicator)	0.0	9.0	1.2	3.5
Address 1 year ago (PC)	1.3	11.2	2.6	3.5

Question	Rehearsal Online (Sumbitted)	Rehearsal Paper	Rehearsal total	2011 total
Workplace/study address (indicator)*	30.4	unavailable	unavailable	8.6
Workplace/study address (PC)*	30.4	unavailable	unavailable	8.5
Method of travel*	30.4	unavailable	unavailable	2.2
Religion	3.4	8.7	4.0	7.0
National identity – tick-box	0.0	3.3	0.5	1.6
Ethnic group – tick-box	0.1	6.7	0.9	2.1
Language skills	N/A	N/A	N/A	1.9
Spoken English proficiency	N/A	N/A	N/A	2.7
English language proficiency	0.1	4.3	0.6	N/A
Gaelic and Scots language skills	0.0	7.7	1.0	N/A
Language at home	0.0	7.4	1.0	3.9
Health	0.0	6.0	0.8	2.3
Long-term condition	0.2	9.3	1.3	15.2
Disability	0.0	6.3	0.8	3.7
Qualifications	0.0	14.6	1.9	6.5
Activity last week	1.7	11.7	3.0	5.6
Ever worked	0.0	11.0	1.4	4.8
Last year worked	N/A	N/A	N/A	16.8
Employee status	0.1	16.6	2.2	4.1
Occupation	0.1	15.5	2.1	4.6
Supervisor status	0.1	16.3	2.1	3.9
Hours worked	0.1	16.6	2.2	4.9
Industry	0.0	19.8	2.5	8.5
Trans status	2.9	10.7	3.9	N/A
Sexual orientation	5.1	20.7	7.1	N/A
British Sign Language	0.0	7.2	0.9	N/A
Ex-service	0.0	9.1	1.2	N/A
Passports	0.0	4.5	0.5	N/A

\* Place of work/study and method of travel for rehearsal paper responses unavailable at this stage due to complexities of applying routing to paper responses. High non-response rates for online responses likely due to known issues around address fields in online questionnaire at time of rehearsal.

\*\* Arrival in UK: Year field used for paper responses, not month. It is easier to separate scanning errors (e.g. “1”, “11”) from year than it is to separate them from genuine responses for month. For online returns, the nonresponse rate is the same for month as for year.