Item editing and imputation process for Census 2021, England and Wales

The methods for resolving item non-response and item inconsistencies in Census 2021 data, including deterministic editing, nearest neighbour donor imputation, and manual imputation.

Contact:
Census Customer Services
census.customerservices@ons.gov.uk
+44 1329 444972

Release date:
8 November 2022

Next release:
To be announced

Table of contents

1 . Introduction and background

Census 2021 was the first digital-first census in England and Wales. There is more information on the digital strategy and results in our Designing a digital-first census article.

As with any self-completion questionnaire, it was possible for respondents to make errors when recording their answers, resulting in data that were not valid for the estimation of population totals. The online questionnaire improved the quality of data collection, when compared with paper response, through validation checks, and eliminated errors encountered when scanning hand-written responses. Online respondents were only presented with questions they needed to answer, which reduced respondent burden and prevented routing errors and some consistency errors. Invalid responses included:

- blanks
- multiple ticks, when only one is required
- out-of-range values
- partially answered responses (for example in occupation, which is collected in multiple fields)

Referred to as "item non-response", this can be unintentional, for example where a respondent misses a question or thinks they can tick more than one option, or intentional, where a respondent does not know the answer or does not want to provide one.

It was common for some correctly recorded values to be considered invalid because they were inconsistent with other values on the questionnaire, or with auxiliary information or definitions. Referred to as "item inconsistency", these errors were detected by validating the data against a set of edit rules.

Edit rules are defined in consultation with users and experts to prevent impossible, highly unlikely, or illegal combinations of values and to ensure data meet expectations. For example, the rule that states that a person aged under 16 years cannot, for census purposes, have a qualification (as the question is only asked of those aged 16 years and over) would have flagged a record where a person gave their age as five and said they had a university degree.

It is important to correct for non-response and inconsistency before using data for further analysis since they can lead to bias and inconsistencies in estimates and analysis. Invalid values can be replaced deterministically using other observed values to deduce a value or through statistical imputation that estimates the unobserved distributions replacing invalid data with plausible data.

Imputation can be implemented to include edit constraints, treating both inconsistencies and non-response simultaneously, or deterministic editing can be used after imputation to correct inconsistencies once non-response has been resolved. Data can also be edited deterministically prior to imputation to correct common response errors. With a few exclusions, editing and imputation were conducted simultaneously for Census 2021 using software that was specifically designed to process census data.

Prior to editing and imputation, several processes were required to turn the paper questionnaire and online responses into data. Details on the processes used prior to editing and imputation will be published in winter 2022. Additional information is currently available showing where edit and imputation fits in the wider Design for Census 2021.

After item editing and imputation, all responding questionnaire records were complete and consistent. However, non-responding or missed persons must be estimated to obtain final census population estimates. The coverage estimation and coverage adjustment processes estimated the missed persons and imputed them into the data, respectively. More information on coverage estimation and adjustment will be published later this year.

Only data that have been edited, imputed, and adjusted are used for official census estimates of population and characteristics. This is because users do not necessarily have all the information required to be able to estimate non-response. It also prevents different methods of editing and imputation being applied that could result in incomparable or contradictory estimates and conclusions being made from the data.

This paper focuses on item editing and imputation for Census 2021 in England and Wales. Similar processes were adopted in Northern Ireland for their Census 2021 and in Scotland for their Census 2022. Any differences in the implementation were made to better reflect the population in each country. Additional information on the UK harmonisation of census outputs is available.

2 . Edit and imputation strategy

The primary objective of the 2021 item editing and imputation (EI) strategy was to produce a fully populated, clean, and consistent unit-level census database by resolving missing values and within- and between-person inconsistencies.

To meet this objective, the EI system was designed according to the following core imputation-specific statistical objectives:

- to use modern statistical EI methodology consistent with internationally recognised principles and standards for large-scale census EI applications, as described in publications included in Section 8

- the methodology should replace missing data and resolve inconsistencies for all responding households and communal establishments with minimal change to the observed data; ensure adjustments to the census database are conditioned by key joint distributions in the data; as far as possible, avoid introducing bias or inconsistency into the data through the imputation process; adjust the database for non-response bias where appropriate; and retain a consistent approach to imputation, minimising the use of alternative fallback methods

- in addition to the editing and imputation of errors in the primary respondent census database, the statistical EI methodology should be available for the Census Coverage Survey and for partial record imputation following the Census Coverage Adjustment, if required

- statistical imputation will always be considered as the first option when there is uncertainty about how an error should be resolved because exceptions could lead to the introduction of bias, so any deterministic rule-based adjustments made should be clearly documented

- the final imputed census dataset should have responses, and relationships between responses, that meet stakeholder expectations of census data with responses and relationships outside of that expectation defined by a set of hard edit rules and resolved by the EI process

- where specified by stakeholders, rare but plausible relationships will be defined by a set of soft ("outlier") edit rules so that the EI process should not propagate rare but plausible relationships between responses in a census record and ensure that the process should also avoid removing observed cases from the census database

It should be noted that item EI does not seek to impute a correct value for a missing or incorrect response. Instead, EI seeks to preserve the observed distribution of a variable, preserve key joint distributions or, where appropriate, adjust the distributions to account for non-response bias. For more information, read the 2021 Census editing and imputation strategy (Word, 1.15MB).

3 . Implementation

Edit and imputation context

In 2021, there were 12 questions about household accommodation, one question about household relationships, and 45 questions about the personal characteristics of each resident. Data consisted primarily of categorical variables governed by implicit and explicit questionnaire routing and a set of consistency edit rules.

The data were hierarchical in the sense that one or more persons were recorded within households. These were imputed jointly, where possible, to preserve the between-person distributions and variance. Many variables were strongly related, both within-person and between-persons. For example, a person's main language, country of birth, and ethnicity were strongly related.

Historically, donor-based methods have been used to impute UK censuses. Donor-based methods are ideal for this type of data because they can handle categorical and numeric variables simultaneously and, when applied correctly, will estimate the distributional properties of the data accurately. There is more information about this in publications included in Section 8.

In donor-based methods, records with invalid values (recipients) are matched to clean records (donors) based on characteristics observed in both records. Value(s) from the donor record are given (donated) to the recipient to make it complete and consistent.

Software

The Canadian Census Edit and Imputation System (CANCEIS) applies nearest-neighbour donor imputation and performs consistency data editing simultaneously. This method selects donors by minimising a specified distance measure between donor and recipient based on auxiliary variables.

Statistical editing is applied by implementing user-defined edit rules that identify combinations of values that are inconsistent. A single donor that offers the minimum number of changes to a record and satisfies edit constraints is used to resolve inconsistencies and non-response.

CANCEIS was evaluated and endorsed as a census imputation tool for Census 2021 in England and Wales following its use in the 2011 Census.

Census 2021 edit constraints

There are two types of edit rules: hard edit rules and soft edit rules. Hard edit rules check the plausibility of data and lead to imputation if the record fails a rule; for example, "a child cannot be older than their parent". Soft edit rules are possible but uncommon values; for example, "a person aged under 16 years is unlikely to be a parent". Soft edits can be applied to prevent records that fail the rule from being used as a donor. This retains the observed records but prevents an increase of these combinations in the database. Alternatively, soft edits may be monitored during editing and imputation to ensure the number of cases that fail the rule is not disproportionately increased. For Census 2021, we used hard edits and both types of soft edit. The edit rules implemented are as follows.

Hard checks

The household hard checks are:

- the number of bedrooms cannot be greater than Valuation Office Agency number of rooms
- if no-one reported living at the address then landlord, number of cars, and ownership type must be "no code required"
- if household size is greater than zero, answers must be provided for ownership and number of cars

Additionally, there is also a dummy hard check, where if a household size is zero, then the reason for the dummy form must be either "second residence", "holiday home", or "vacant".

The within-person hard checks are:

- if aged under 16 years, then marital and civil partnership status cannot be in a civil partnership, separated but still legally in a partnership, formerly in a civil partnership which is now legally dissolved, or surviving partner from civil partnership
- a person aged between 6 and 15 years old must be a student in full-time education unless limited a lot by a health problem or disability
- a person cannot arrive to live in the UK before their date of birth
- a person cannot be aged under 17 years old and usually travel to work driving a car or van
- a person who has never worked and is not currently working cannot have a second address reason of "another address when working away from home"
- a person intending to stay less than 12 months in the UK cannot have arrived before March 2020
- a person who is not working cannot have a position of staff in a communal establishment
- if second address is the same as workplace address, second address type should include "another address when working away from home" or "armed forces base address"
- if second address type is "student's term-time address", then term-time address cannot be "address on front of questionnaire"
- a person who has a "method of travel to work" value of "working from home" cannot have a workplace address value other than "working from home"

Hard checks are also applied to the relationships between people. These checks are:

- a person cannot have more than one spouse or civil partner
- a person cannot have a spouse or civil partner and a partner
- two people with at least one parent in common cannot be married, civil partners or partners with each other
- a person aged under 16 years cannot be a civil partner
- a parent cannot be less than 12 years older than their child
- a person aged under 15 years cannot be a spouse
- a person aged under 12 years cannot be a partner or a stepparent
- 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
- a person cannot have more than two parents and two stepparents
- a grandparent cannot be less than 24 years older than their grandchild
- a woman cannot be more than 66 years older than their child
- two people with a parent in common must be siblings
- if two people are siblings and one has a parent, the other sibling must be a child or stepchild to their sibling's parent

Soft checks

Soft checks can be applied. The applied soft checks applied are:

- a father is unlikely to be more than 65 years older than his child
- a person aged under 16 years is unlikely to be a partner
- a person aged 42 years and over is unlikely to be a student in full-time education
- a person is unlikely to have a marital and civil partnership status of civil partner of the opposite sex (includes separated, dissolved, and surviving)

The soft check rule of "a person is unlikely to have a marital and civil partnership status of civil partner of the opposite sex (includes separated, dissolved, and surviving)" was introduced during processing to prevent the use of records who met the condition as donors. This is because their marital status was, in many cases, incorrectly recorded. The number of observed opposite-sex civil partners were far greater than the number of legally registered opposite-sex civil partners. Similarly, there were many observed records with dissolved opposite-sex civil partners, yet no registered opposite-sex civil partnership had been dissolved as of Census Day.

Soft household checks are also monitored. These were:

- if household accommodation is in purpose-built flats or tenement, part of a converted or shared house, in a commercial building, or in a caravan or other mobile or temporary structure, then number of usual residents is unlikely to be more than ten
- houses or bungalows that are detached, semi-detached, or terraced and flats, and maisonettes or apartments that are in a purpose-built block of flats or tenement, are likely to be self-contained
- a whole house or bungalow is unlikely to have fewer than two rooms
- a caravan or mobile home or temporary structure is unlikely to have more than eight rooms
- a caravan or mobile home or temporary structure is unlikely to have central heating
- it is unlikely that there will be more than two people for each room
- it is unlikely that there will be more than three people for each bedroom
- accommodation rented from a council, local authority, private landlord, or letting agency is unlikely to be rent free

The within-person soft checks that are monitored are:

- a person aged under 28 years is unlikely to have a marital and civil partnership status of divorced or legally dissolved same-sex civil partnership
- the address one year ago is unlikely to be in the UK if date of arrival in the UK is after March 2020; this is for England and Wales only
- a person with an apprenticeship is unlikely to never have worked
- it is unlikely for someone aged under 41 years to have a marital and civil partnership status of widowed
- a person aged 71 years and over is unlikely to have an economic activity status last week of "working"
- a person aged under 16 years is unlikely to have a second address type of "Armed forces base address"
- a person aged 65 years and over is unlikely to have a second address type of "Another parent or guardian's address"
- a person aged 15 years with country of birth "elsewhere" is unlikely to have a marital and civil partnership status of married or separated but still legally married, divorced, or widowed
- it is unlikely for someone aged under 41 years to have a marital and civil partnership status of "surviving partner of civil partner"
- a person aged under 4 years or 65 years and over is unlikely to have the reason for second address as "student's home address"
- a person aged under 16 years is unlikely to have a reason for second address as "another address when working away from home"
- a person is unlikely to describe their national identity as Welsh if they have ticked "none of the above" for Welsh language proficiency and do not speak English "well" or "very well" (unless aged under 4 years or have British Sign Language (BSL) as their main language)
- a person is unlikely to describe their national identity as English if they cannot speak English well or very well (unless aged under 4 years or have BSL as their main language)
- a person aged under 6 years or 65 years and over is unlikely to have their address one year ago as "student term-time or boarding school address"
- a person aged under 54 years is unlikely to have "economic activity status last week" as "retired"
- a person aged 65 years and over is unlikely to have "economic activity status last week" as "student"
- a resident in a communal establishment with establishment group type of "elderly care" is unlikely to be aged under 21 years
- a resident in a communal establishment with establishment group type of "education" is unlikely to be aged 41 years and over
- a resident in a communal establishment with establishment group type of "armed forces" is unlikely to be 48 years and over
- a resident in a communal establishment with establishment group type of "detention" is unlikely to be aged 72 years and over

Note that the establishment group type variable was used during processing but was not edited or imputed.

Census 2021 also monitored other soft checks between people in households. These include:

- a person with a parent aged under 30 years in the household is unlikely to have a marital status other than single
- a person with a grandparent aged under 40 years is unlikely to have a marital status of married or separated, but still legally married, divorced, widowed, in a registered same-sex civil partnership, or separated but still in a registered same-sex civil partnership
- it is unlikely that at least one usual resident will not be aged 16 years and over
- a person aged under 16 years is unlikely to be a husband or wife
- a person aged under 16 years is unlikely to be a stepparent
- a person aged under 14 years is unlikely to be a mother or father
- a parent is unlikely to be less than 14 years older than their child
- a mother is unlikely to be more than 49 years older than her child
- siblings are unlikely to have an age difference of more than 37 years
- a stepchild is unlikely to be older than a stepparent

The strategy for Census 2021 was to edit and impute simultaneously wherever possible. There were 32 edit rules applied during imputation. These were based on the 2011 edit rules but updated through consultation with users and subject matter experts. For example, the rule that in 2011 said married couples had to be of the opposite sex and civil partners had to be the same sex was removed entirely following changes to legislation that allowed for both opposite and same sex marriages and civil partnerships.

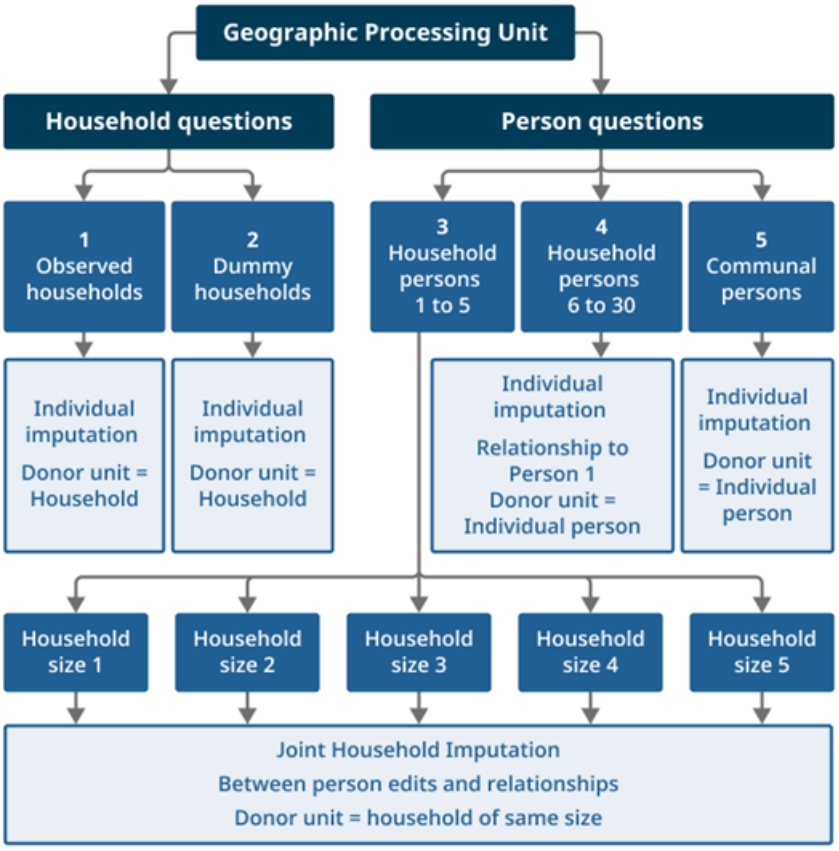
Imputation groups

With a base population of 58.6 million people, it was not possible to impute the whole database in one pass through CANCEIS because of system resource constraints. Additionally, not all variables were imputed simultaneously because this would have reduced donor levels below a viable level, compromising data quality (one invalid value for one person would prevent their whole household from being used as a donor).

To overcome these two issues, the data were partitioned into manageable imputation groups, illustrated in Figure 1. Firstly, the data were split into 101 geographical regions, or processing units (PUs), containing on average 261,000 households with 572,000 person records. The data in each PU were split into "household questions", where there was only one response for each household, and "person questions" where there was a separate response for each person.

Observed households were imputed together, as were all the dummy households (where an enumerator completed limited questions through observation because of a household not returning a questionnaire), giving the first two imputation groups. The person data were divided into three groups. Firstly, persons one through five collected on the main household questionnaire were imputed jointly with other households of the same size. Secondly, persons 6 to 30 collected on household continuation questionnaires were imputed together as individuals. Lastly, persons from communal establishments collected on individual questionnaires were imputed as individuals. The five broad groups for imputation labelled numerically in Figure 1 correspond to the data groups shown in Figure 2.

Figure 1: Data groups and imputation types within each processing unit

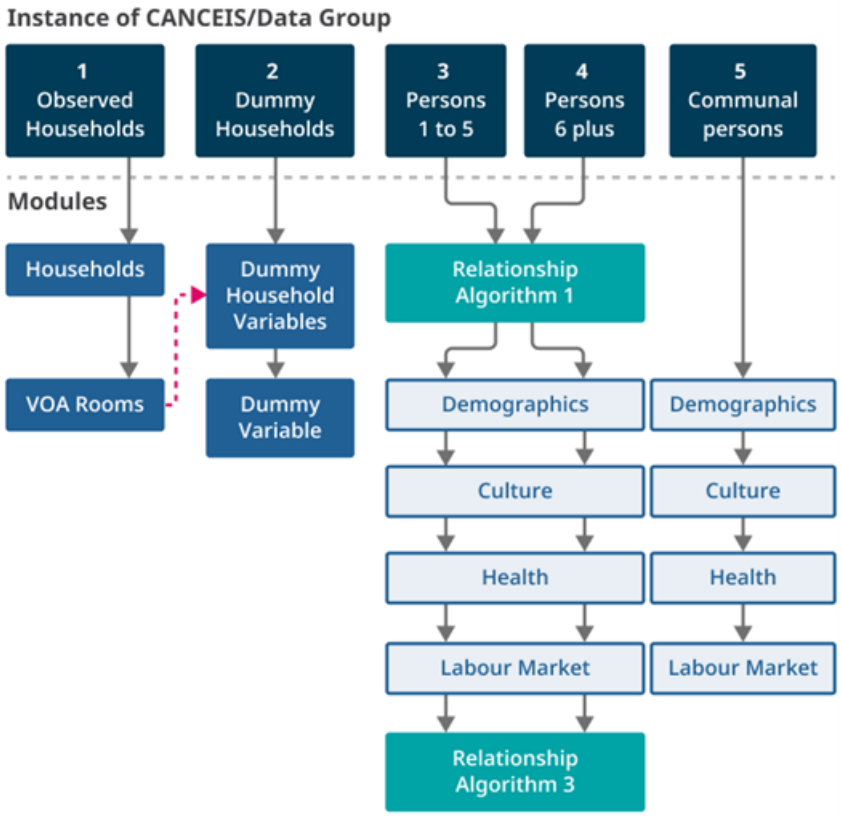


Source: Office for National Statistics

Each data group consisted of two or more modules, in which groups of variables were imputed simultaneously using CANCEIS. There were two main aims in forming the imputation modules: to have variables that help to predict each other and to maximise the number of donors for a given group. Other factors also influence their structure, including the order of the questionnaire and its routing, the priority of each variable, the inter-relatedness of the variables, and the edit rules.

The first seven imputable household variables were imputed in the first module in Figure 2. In 2021, no question was asked about the total number of rooms in a household. Instead, numbers of rooms was joined onto census data from an alternative administrative data source (the Valuation Office Agency (VOA)). This was imputed in the VOA rooms module. Next, the subset of household questions collected for dummy households was imputed using clean records from the household module as donors. The second dummy household module imputed only one question, "reason for dummy return", using otherwise complete and consistent dummy households as donors.

Figure 2: Implementation of imputation groups in CANCEIS



Source: Office for National Statistics

Notes:

1. CANCEIS refers to the Canadian Census Edit and Imputation System.

The large number of person variables (42) was divided into four modules for imputation. The first group consisted of the main "demographics" variables, such as:

- age
- sex
- economic activity status last week
- marital and civil partnership status

The second "culture" module contained additional details about the respondents' background, for example:

- ethnicity
- country of birth
- main language

The third group contained the health-related and caring responsibility questions and the final group, "labour market", contained questions on employment and qualifications. In every PU, there were records from both online and paper modes. CANCEIS was set up to only select donors from a different mode if there were no suitable donors in the same mode as the recipient.

Responses to the voluntary questions on gender identity, religion, and sexual orientation were not imputed during main item edit and imputation; responses, including non-response, remained unchanged throughout. These, voluntary questions were not used as matching variables to impute any other variable, nor were they used for any other process including deterministic editing or fallback methods.

The dotted line in Figure 2 illustrates where information was shared between components. Here, the clean observed household records from the households module were passed into the first dummy household module (dummy household variables). Figure 2 also illustrates that persons one to five and persons six and over simultaneously passed through Relationship Algorithm 1, which deterministically edited common errors in the relationship questions. Processing data simultaneously helped to maintain consistency within households in the relationship data despite these groups being processed separately.

Persons one to five were passed through their respective modules, followed by persons six and over through theirs. All persons in households were finally passed through Relationship Algorithm 3 to impute relationships that could not be treated by CANCEIS and to ensure consistency between relationships in households where CANCEIS had been applied (see Section 4 for more information on the relationship algorithms). Finally, persons in communal establishments were processed in CANCEIS through their modules.

Within each module, a model specifies matching variables and distance functions used in the selection of potential donors. The overarching aim of imputation is to improve the utility of the data, and the main analytical aims of the survey were factored into the design of the imputation process. To this end, matching variables were largely based on categories used for the Census 2021 outputs. This helped to estimate the unobserved marginal distributions for important outputs.

Each matching variable was weighted according to several factors, including how well they were expected to predict other values and how highly they should be prioritised when resolving inconsistencies. For example, age is often a good predictor of other demographic variables. Therefore, age was given a high weight in the model and observed ages were prioritised over other values if there was an inconsistency.

Fallback methods

CANCEIS has many parameters that allow users to specify how it operates including how many records to search and how similar records must be to be considered a potential donor. Every PU first passed through an automated process with an optimal set of parameter values. A small number of records in each PU failed to impute first time. In all cases, these were household-person questions. No household questions failed to find donors and only a single communal establishment-person record failed on its first run.

Since statistical imputation was always preferred, the system was set up to detect failures and pass failed records back through the automated system. On this second pass, the potential donor pool included the extra repaired records from the first pass, plus CANCEIS was also given a set of relaxed parameter values. The relaxed parameter values enabled CANCEIS to look for donors both statistically and geographically further away from the recipient.

Out of 58.6 million records, there were 1,316 records that could not be imputed after a second automated attempt. For these, one or more values were manually edited either to completely resolve a record or, more often, to enable the system to find suitable donors.

The final step in the automated process was Relationship Algorithm 3 (RA3). RA3 was only able to assess relationships between three people at a time. Occasionally, it could not determine an appropriate set of consistent relationships within a (typically large) household. Similarly, issues were encountered when there was high non-response in relationship values. In these cases, the demographic variables of households were inspected to deduce consistent relationship values. Such instances formed the vast majority of manual edits.

In all cases, when manually editing records, only the minimum number of changes necessary was made to ensure the automated system could pass all records. Even in cases of a single item failure for a single record, the entire PU was passed back through the whole system to ensure records' adherence to the edit rules and consistency checks.

4 . Relationship algorithms and deterministic edits

The relationship question was a matrix-style question that collected the relationships between persons recorded on the questionnaire. The relationship question can be viewed on page four and page five of the Census 2021 paper questionnaires for households.

For paper respondents, large households required a continuation questionnaire(s). Additional (continuation) persons were only linked to "person one" on the main questionnaire and to each other within the same form. The same relationship information was collected by the online questionnaire.

It was anticipated that more response errors would be observed for the relationship question, especially for paper response, since this was seen in 2011. The relationship question needed to be imputed in the demographics module together with other priority variables like age, sex, and marital and civil partnership status to maintain the edit rules. Higher levels of inconsistency or non-response in relationships could affect the quality of the imputation for important demographics by reducing the donor pool and possibly undermining minority sub-populations.

In 2011, a selection of deterministic edits, where the correct responses could be deduced with a high level of certainty, was developed. This was adapted for use in 2021 and applied prior to imputation to improve the quality of relationship variables. In Figure 2, these rules relate to Relationship Algorithm 1 and are as follows.

Relationship Algorithm 1 conditions

Condition 1: child is 13 years or more older than their parent

If the child is aged under 30 years and living at home, then:

- the action will be to change relationship from parent to child or vice versa
- the constraints will be that the person to be made a child does not have a partner in the house, has a marital and civil partner status of single, and is not aged 30 years and over, and the person to be made a parent is aged 16 years and over

Condition 2: stepchild is 5 years or more older than their stepparent

If the stepchild is 5 years or more older than their stepparent, then:

- the action will be to change the relationship from stepparents to stepchildren or vice versa
- the constraints will be that the person to be made a stepchild does not have a partner in the house, has a marital and civil partner status of single, is not aged 30 years and over, is aged under 19 years, and the person to be made a stepparent has a valid spouse or partner relationship with the other parent of the person to be made a child

If the stepchild is aged 40 years and over living with lone parent, then there will be no action, as we cannot valid the stepparent relationship because there is no partner in the house, and constraints will be non-applicable.

Condition 3: child is 26 years or more older than their grandparent

If the grandchild is 26 years or more older than their grandparent, then:

- the action will be to change the relationship from grandparent to grandchild or vice versa
- the constraints will be that the person to be made a grandchild does not have a partner in the house, has a marital or civil partner status of single, and the person to be made a grandparent is aged 32 years and over

Condition 4: A person with a grandchild is not recorded as a grandparent or a person with a grandparent is not recorded as a grandchild

If a person with a grandchild is not recorded as a grandparent, or a person with a grandparent is not recorded as a grandchild, then:

- the action will be to assign a grandparental relationship
- the constraints will be that the grandparent is a least 13 years older than their child, the parent is at least 16 years older than the child, the person to be made a grandchild does not have a partner in the house, has a martial and civil partner status of single, and is aged under 30 years

Condition 5: Two persons sharing one or more parents and not recorded as siblings or stepsiblings

If siblings of a lone parent, then:

- the action will be to assign a sibling relationship
- the constraints will be that the parent is at least 13 years older than both children and aged 16 years and over, and the siblings are not more than 20 years different in age

If siblings of two parents, then:

- the action will be to assign a sibling relationship
- the constraints will be that the parents are both at least 13 years older than both children and are both aged 16 years and over, and that the siblings are not more than 20 years different in age

If siblings of a parent and stepparent, then:

- the action will be to assign a stepsibling relationship
- the constraints will be that the parent is a least 13 years older than both children, the stepparent is at least 5 years older than both children, the parent is aged 16 years and over, the step-parent is aged 18 years and over, and the siblings are not more than 20 years different in age

Condition 6: A person has a spouse and another member of family in the house and these two persons are not recorded as "other relation"

If they are a spouse and a parent, then:

- the action will be to assign an "other relation" relationship
- the constraints will be that the relationship is currently recorded as missing or unrelated, the parent is at least 13 years older than the child, the child and their spouse or civil partner or partner are both aged 18 years and over, and the siblings are not more than 20 years different in age

Condition 7: A person with two siblings or stepsiblings in the house and these two persons are not reported as siblings

If a person has two siblings or stepsiblings in the house and these two persons are not reported as siblings, then:

- the action is to assign a sibling or stepsibling relationship
- the constraints will be that the relationship is currently recorded as unrelated, and all siblings are not more than 20 years different in age

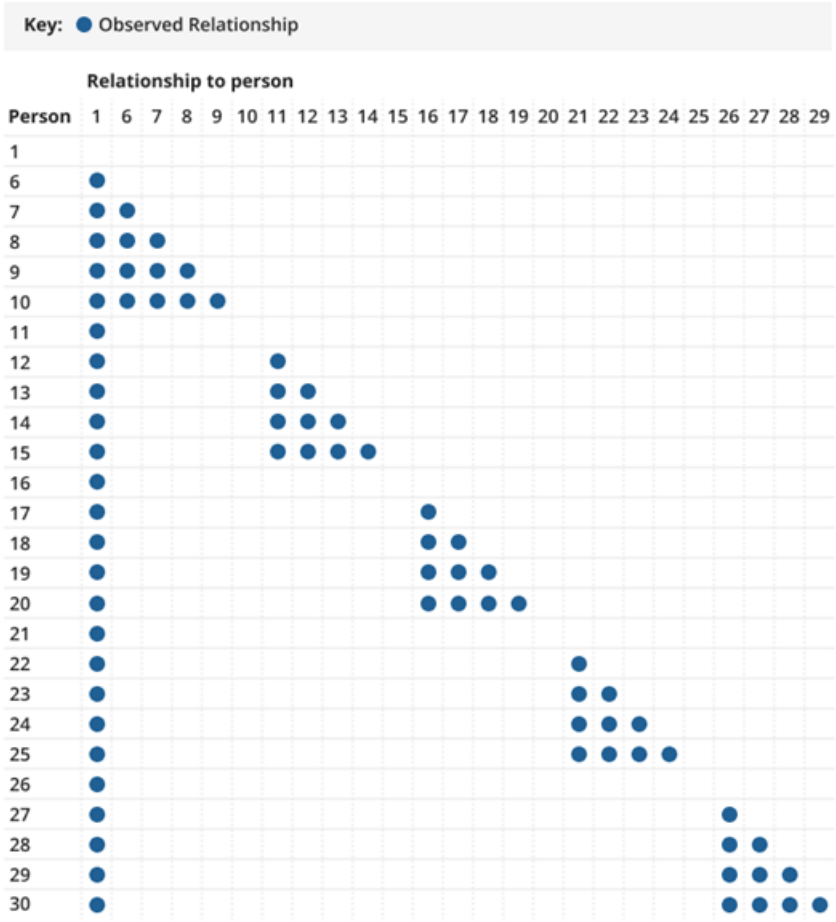
Condition 8: A parent of a child in the house has been miscoded as a sibling to the child, where the parents have a valid partner relationship

If a parent of a child in the house has been miscoded as sibling to the child, where the parents have a valid partner relationship, then:

- the action is to assign a parental relationship
- the constraints will be that the person to be made a parent has a partner in the house and is at least 13 years older than the suspected miscoded sibling and must have a valid parental relationship to the suspected miscoded sibling, must have a valid parental relationship to at least one other child in the house, and the two children are recorded as siblings with an age difference of less than 20 years

Similarly, Relationship Algorithm 3 (RA3) was developed in 2011 to impute the relationships of continuation persons not treated by the Canadian Census Edit and Imputation System (CANCEIS) and ensure consistency after imputation. RA3 was also adapted and implemented for use in 2021. The observed relationships for persons six and over are shown in Figure 3; only those in the first column were treated by CANCEIS, and the remainder were imputed by RA3.

Figure 3: Observed relationship for persons six and over



Source: Office for National Statistics

Notes:

1. CANCEIS refers to the Canadian Census Edit and Imputation System.

The adaptations for 2021 were to account for two main changes. The questionnaire structure changed between 2011 and 2021 (five persons were collected on the main form in 2021 versus six in 2011, resulting in changes to all continuation forms). Each algorithm was also updated to reflect changes to edit rules such as the inclusion of both opposite and same sex married couples and civil partners.

There were no other deterministic edits implemented as part of the 2021 main-item edit and imputation process.

5 . Comparing 2011 and 2021

The implementation of item-level editing and imputation (EI) in 2021 was very similar to that in 2011. There were, however, a few notable improvements and differences introduced or influencing the processes.

Online response made up a small percentage of overall response in 2011 and response mode was not explicitly a factor during imputation. Since the majority of responses were online in 2021, the electronic questionnaire, with its automatic routing and verification of responses, reduced respondent burden and thus inconsistencies and item-non-response when compared with 2011 (see Table 1 for further information). In 2021, online responses were repaired primarily using donors who also responded online. Similarly, paper responses were primarily repaired using donors who responded on paper. The matching of donors and recipients by mode reduced the risk of introducing bias into these broadly distinct populations. There is more information available in Rogers, Dyer and Foley's 2014 publication Towards the 2021 UK Census imputation strategy: Response mode as a matching variable in a donor-based approach?

One of the processes used prior to imputation was the "student-method" (more information to be published in January 2023), which corrected for student displacement because of the coronavirus (COVID-19) pandemic. Item-level responses for affected individuals could not be copied across databases, resulting in an increase of apparent non-response for most non-demographic variables for these records. This was, however, in the context of lower overall non-response compared with 2011, and the system proved resilient to this.

Relationship Algorithm 2 (RA2) was deployed in 2011 before the imputation of persons seven and over. An evaluation found that it only treated a minority of records but, since all the algorithms were complex, it had a high processing time. As such, RA2 was not considered critical and was dropped for processing in 2021 without loss of quality.

In 2011, where a record was edited deterministically but the value was changed again by the Canadian Census Edit and Imputation System (CANCEIS), it contributed to both the deterministic imputation rate and the item imputation rate. For 2021, we do not double count imputations. Imputation rates are defined and reported according to the imputation that occurred last. For example, for a variable with both Relationship Algorithm 1 and CANCEIS edits, only the CANCEIS edits contribute to the imputation counts or rates.

Finally, for the first time in 2021, the use of linked administrative data successfully replaced the collection of the "number of rooms" household question with Valuation Office Agency data on number of rooms.

Table 1 compares operational results from 2011 and 2021. Note that population figures may differ from official estimates.

Table 1: Operational comparison of main-item edit and imputation in 2011 and 2021, England and Wales
England and Wales, 2021

	Census 2011	Census 2021
Persons		
Person records processed	53.5 million	58.6 million
Average number of records in a PU	530,000	580,000
Average time to impute a PU	12 hours	5 hours
Persons needing at least one question imputed	18.6 million (35%)	15.6 million (26%)
Percent imputed as a household taking account of joint distributions between questions	82%	93%
Percent imputed as individuals	18%	7%
Percent imputed using fallback methods	0.10%	<0.003%
Maximum donor reuse allowed	100	10
Households		
Household records processed	24.3 million	26.3 million
Households requiring at least one item imputed.	2.8 million (9.5%)	2.4 million (9%)
Percent imputed taking into account multivariate join distribution between questions	100%	100%
Maximum donor reuse allowed	50	50

Source: Office for National Statistics

Overall, in 2021, a complete and consistent database was achieved, and item editing and imputation was successful in meeting the main objectives and aims outlined in the strategy. More records were processed in a faster time, with a greater proportion of imputation taking account of joint distributions. There was less reliance on fallback methods and maximum donor reuse was lowered where possible, reducing the risk of introducing bias during imputation through characteristics from a single donor being imputed into many recipients.

6 . Item non-response, edit failure, and imputation rates

Overall imputation rates

Table 1 shows there were 26.3 million household returns in 2021 and 2.4 million (9%) of these required one or more questions to be imputed. All household questions were imputed by joint imputation with a single donor. There were also 58.6 million person returns with 15.6 million (26%) requiring one or more questions to be imputed, of which, 93% of persons were imputed by joint imputation.

The household module

Item non-response (Table 2) ranged from 0.8% (self-contained) to 2.5% (landlord type). While missingness in the administrative data for Valuation Office Agency number of rooms was higher at 3.2%. There were three edit rules for the household questions and a questionnaire filter between tenure of household and landlord type. Less than 0.2% of observations for accommodation type, type of heating in household, number of bedrooms, car or van availability, ownership type, and self-contained were changed because of failing the edit rules or the filter.

Table 2: Household module: non-response, edit failure and imputation counts and rates
England and Wales, 2021

	All households	Non-response	Edit failure	Total imputed	Non-response	Edit failure	Total imputed
	N	N	N	N	%	%	%
Accommodation type	24,817	227	0	227	0.9	0	0.9
Type of central heating in household	24,817	535	0	535	2.2	0	2.2
Landlord type	9,051	227	50	277	2.5	0.5	3.1
Number of bedrooms	24,817	268	0	269	1.1	0	1.1
Car or van availability	24,119	230	26	256	1	0.1	1.1
Tenure of household	24,119	243	29	272	1	0.1	1.1
Self-contained	24,817	188	0	188	0.8	0	0.8
Number of rooms (Valuation Office Agency)	24,817	787	336	1,123	3.2	1.4	4.5

Source: Office for National Statistics

Notes

1. All counts are expressed in thousands.

2. Counts and percentages cover all responding households eligible to answer each question, excluding dummy household returns.

3. Car or van availability, tenure of household, and landlord type only answered by households with at least one resident.

4. Landlord type includes households with at least one resident and tenure of household is renting, part renting, or rent free.

Demographic module

The demographics module included personal characteristics such as age and sex, marital and civil partnership status, and relationships. All questions up to the student term-time question were answered by every person; students with another address during term time were counted at both their home and term-time addresses for all questions except economic activity status last week.

Table 3: Demographic modules: non-response, edit failure and imputation counts and rates
England and Wales, 2021

	All Persons	Non-response	Edit failure	Total imputed	Non-response	Edit failure	Total imputed
	N	N	N	N	%	%	%
Age	58,624	89	32	121	0.2	0.1	0.2
Sex	58,624	160	9	169	0.3	0	0.3
Marital and civil partner status	48,464	398	559	957	0.8	1.2	2
Second address indicator	58,624	411	70	481	0.7	0.1	0.8
Second address country	791	8	5	12	1	0.6	1.5
Second address postcode	2,909	79	3	82	2.7	0.1	2.8
Second address type	3,700	29	43	72	0.8	1.2	2
Term-time address indicator	11,832	22	75	98	0.2	0.6	0.8
Schoolchild or full-time student indicator	55,551	299	131	429	0.5	0.2	0.8
Economic activity status last week	47,205	367	18	385	0.8	0	0.8
Relationship to person one	33,646	503	284	787	1.5	0.8	2.3
Position in communal establishment	859	157	13	170	18.3	1.5	19.8
Gender identity	47,205	2,786	-	-	5.9	-	-
Sexual orientation	47,205	3,482	-	-	7.4	-	-

Source: Office for National Statistics

Notes

1. All counts are expressed in thousands.

2. Counts and percentages cover all responding persons eligible to answer each question.

3. Age, sex, and second address indicator were answered by all persons, including students at their home (non-term-time) address.

4. Marital and civil partner status was answered by all persons aged 15 years and over, including students at their home (non-term-time) address.

5. Schoolchild or full-time student indicator was answered by all persons aged 5 years and over, including students at their home (non-term-time) address.

6. Term-time address indicator was answered by all persons aged 5 years and over responding “yes” to the schoolchild or full-time student indicator question.

7. Economic activity status last week was answered by all persons aged 16 years and over, not including students at their home (non-term-time) address.

8. Relationship to person one was answered by persons 2 to 30 living in households of size two or greater, including students at their home (non-term-time) address.

9. Position in communal establishment was only answered by people living in communal establishments. 10. Gender identity and sexual orientation were voluntary questions answered by all persons aged 16 years and over, not including students at their home (non-term-time) address.

10. Gender identity and sexual orientation did not pass through the edit and imputation system.

Table 3 shows the highest level of item non-response in compulsory demographic questions for persons was in second address postcode (2.7%). Position in communal establishment was only asked of persons in communal establishments and non-response was very high at 18.3%. Non-response for the important variables age and sex was very low at 0.2% and 0.3%, respectively.

This module had the most edit rules (22), as well as six questionnaire filters. However, imputation because of inconsistency or edit failure was mostly low. The highest edit rule failure rate was the communal establishment variable (1.5%), followed by marital and civil partnership status (1.2%)

Culture module

The culture module included questions about characteristics such as country of birth, ethnicity, and main language. These were answered by all persons except students who had another address during term time.

Table 4: Culture modules: non-response, edit failure and imputation counts and rates
England and Wales, 2021

	All Persons	Non-response	Edit failure	Total imputed	Non-response	Edit failure	Total imputed
	N	N	N	N	%	%	%
Country of birth	57,924	258	46	304	0.4	0.1	0.5
Number of months since arrival in UK	9,585	287	8	296	3	0.1	3.1
Intention to stay in the UK	618	20	0	20	3.2	0.1	3.2
National identity UK	57,834	416	83	499	0.7	0.1	0.9
National Identity rest of world	6,822	138	3	140	2	0	2.1
Welsh language proficiency	2,916	42	0	42	1.4	0	1.4
Main language	56,157	521	9	530	0.9	0	0.9
Proficiency in English language	4,903	50	1	51	1	0	1
Address one year ago indicator	57,370	796	249	1,045	1.4	0.4	1.8
Address one year ago postcode	5,645	160	20	179	2.8	0.3	3.2
Address one year ago country	585	6	11	17	1.1	1.9	3
Passports held	57,924	585	58	642	1	0.1	1.1
Passports held (non-UK)	6,313	152	2	154	2.4	0	2.4
Ethnic group	57,924	757	5	762	1.3	0	1.3
Religion	57,924	3,456	-	-	6	-	-

Source: Office for National Statistics

Notes

1. All counts are expressed in thousands.

2. Counts and percentages cover all responding persons eligible to answer each question.

3. All questions eligible to be answered by all persons except students at their non-term-time address.

4. Number of months since arrival in UK only answered by persons where country of birth was outside the UK.

5. Intention to stay in the UK only answered by persons where country of birth was outside the UK and where arrival to the UK was less than 12 months ago.

6. National identity “rest of world” only answered by persons that have at least one non-UK national identity. 7. Welsh language proficiency only answered by persons where country of residence is Wales.

7. Proficiency in English language only answered by persons where main language is not English (or Welsh in Wales) and age is greater than two years.

8. Address one year ago variables only answered by persons aged one year and over.

9. Address one year ago postcode only answered by persons with a different address in the UK one year prior to Census Day.

10. Address one year ago country only answered by persons with a different address outside of the UK one year prior to Census Day.

11. Passport held (non-UK) only answered by persons with at least one non-UK passport.

12. Religion was a voluntary question that did not pass through the edit and imputation system.

Table 4 shows non-response varied from 0.4% (country of birth) to 3.2% (intention to stay in the UK), though for more than half of variables non-response was low at below 1.5% (note that religion was a voluntary question). There were six filter rules and two edit rules, and generally the edit failure rate was very low with only two variables greater than 0.5%.

Health module

The health module treated the three questions related to health and caring responsibilities. A response was required for all persons except students who had another address during term time. Also, the unpaid carer question was only answered by persons aged 5 years and over. Table 4 shows non-response was less than 1.4% across all questions. With only one edit rule and one filter, the edit failure rates were very low at less than 0.1%.

Table 5: Health modules: non-response, edit failure and imputation counts and rates
England and Wales, 2021

	All Persons	Non-response	Edit failure	Total imputed	Non-response	Edit failure	Total imputed
	N	N	N	N	%	%	%
Disability	57,924	751	3	754	1.3	0	1.3
General health	57,924	487	3	490	0.8	0	0.8
Unpaid care	54,851	639	3	642	1.2	0	1.2

Source: Office for National Statistics

Notes

1. All counts are expressed in thousands.
2. Counts and percentages cover all responding persons eligible to answer each question.
3. All questions eligible to be answered by all persons except students at their non-term-time address.
4. Unpaid care only answered by persons aged 5 years and over.

Labour market module

The labour market module contained all questions about qualifications and working, and these were answered by all persons aged 16 years and over, except students who had another address during term time. However, hours worked, method used to travel to work, workplace type, workplace address postcode, and workplace address country could only be answered if respondents had a job in the week prior to Census Day.

Table 6: Labour market modules: non-response, edit failure and imputation counts and rates
England and Wales, 2021

	All Persons	Non-response	Edit failure	Total imputed	Non-response	Edit failure	Total imputed
	N	N	N	N	%	%	%
Qualifications	47,205	1,259	6	1,265	2.7	0	2.7
Has ever worked	20,101	401	96	497	2	0.5	2.5
Employment status	42,146	686	29	715	1.6	0.1	1.7
Occupation (former)	15,042	788	0	788	5.2	0	5.2
Occupation (current)	27,104	822	29	851	3	0.1	3.1
Industry (former)	15,042	2,114	0	2,114	14.1	0	14.1
Industry (current)	27,104	2,704	29	2,733	10	0.1	10.1
Supervisor of employees	42,146	1,004	29	1,034	2.4	0.1	2.5
Method used to travel to work	27,104	307	1,560	1,867	1.1	5.8	6.9
UK armed forces veteran indicator	47,205	1,056	5	1,061	2.2	0	2.2
Workplace address country	117	17	4	21	14.6	3.7	18.3
Workplace address postcode	14,726	2,518	603	3,121	17.1	4.1	21.2
Workplace type	27,104	351	2,353	2,704	1.3	8.7	10
Hours worked	27,104	365	69	434	1.3	0.3	1.6

Source: Office for National Statistics

Notes

1. All counts are expressed in thousands.

2. Counts and percentages cover all responding persons eligible to answer each question.

3. All questions eligible to be answered by all persons aged 16 years and over, except students at their non-term-time address.

4. Has ever worked only answered by persons not currently working as an employee, not self-employed or freelance, not temporarily away from work, or not on maternity or paternity leave.

5. Employment status and supervisor of employees is only answered by persons who are currently working or have previously worked.

6. Occupation (former) and industry (former) only relate to persons who are not currently working but have previously worked.

7. Occupation (current) and industry (current) only relate to persons who are currently working.

8. Method used to travel to work, workplace type, and hours worked only answered by persons employed in the week prior to Census Day or who were temporarily away from work.

9. Workplace address postcode and workplace address country only answered if workplace type is at a workplace or depot in the UK or a workplace or depot in another country.

Table 6 shows non-response was notably higher for industry, workplace address country, and workplace address postcode compared with other variables in this module, ranging from 10.0% to 17.1%. This level of non-response was consistent with that observed in 2011. With five filter rules and four edit rules, edit failure rates were low in general, but transport to work and workplace type were notably higher at 5.8% and 8.7%, respectively.

7 . Further information

Topic summary reports covering each Census 2021 question will be available from the Census 2021 website in future releases. Also, the item non-response, editing and imputation rates will be available to download for countries, regions, counties, unitary authorities, and local authorities.

8 . Related links

A systematic approach to automatic edit and imputation
Article | Released March 1976
This article is concerned with the automatic editing and imputation of survey data.

Handbook of Statistical Data Editing and Imputation (PDF, 199KB)
PDF | Released March 2011
A practical, one-stop reference on the theory and applications of statistical data editing and imputation techniques.

Extending the Fellegi-Holt Model of Statistical Data Editing (PDF, 63.6KB)
PDF | Released February 2002
This paper provides extensions to the theory and the computational aspects of the Fellegi-Holt Model of Editing (JASA 1976).

Multiple Imputation for Nonresponse in Surveys
Book | Released June 1987
Demonstrates how nonresponse in sample surveys and censuses can be handled by replacing each missing value with two or more multiple imputations.

Nearest Neighbour Imputation for Survey Data (PDF, 156KB)
PDF | Released March 1999
Nearest neighbour imputation is one of the hot deck methods used to compensate for nonresponse in sample surveys.

Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review (PDF, 321KB)
PDF | Released June 2005
Nonresponse is a major problem often faced by social scientists when analysing survey data. A range exists to impute the missing responses but the choice between these methods may be difficult. This paper reviews advantages and disadvantages of a range of imputation methods and provides guidance on how to use such methods in practice.

9 . Cite this methodology

Office for National Statistics (ONS), released 8 November 2022, ONS website, methodology, Item editing and imputation process for Census 2021, England and Wales