Improving the accuracy of the Census 2021 data by resolving multiple responses (RMR)

Methodology for resolving multiple responses in Census 2021 data.

Contact:
Census customer services
census.customerservices@ons.gov.uk
+44 1392 444972

Release date:
22 June 2023

Next release:
To be announced

Table of contents

1 . Background

One of the census programme's primary aims was to develop a statistical census database that can be used by government, academics, and other researchers to develop a rich and wide range of population estimates. Accurate population estimates are vital for ongoing planning and development of social and financial policy. However, although a considerable amount of effort and resources are invested in the census collection strategy the raw data will inevitably contain errors and inconsistencies. Consequently, it has to be cleaned, adjusted, and quality assured through a series of deterministic and statistical methods carefully designed to improve its accuracy and utility.

In this paper, we outline and evaluate one of the main preliminary census data-cleaning methods, Resolve Multiple Responses (RMR).

2 . The role of Resolve Multiple Responses (RMR)

A widely held expectation of the Census 2021 database is that it will contain a unique record of every communal establishment (CE) and household (HH) in England and Wales, and, in cases where the property is occupied, an equally unique record of every resident. However, as with previous censuses, meeting this expectation is not always achieved through receipt of a single completed census questionnaire collected from a given address. Quite often, and for many reasons, a respondent will provide information about themselves and where they live more than once. These are referred to as multiple or duplicate responses.

It is worth noting here that RMR assumes that a CE is a property with full-time or part-time supervision providing residential accommodation, such as student halls of residence, boarding schools, armed forces bases, hospitals, care homes, and prisons. An HH is any other residential property where someone might live, such as a house, a flat, a bungalow, a caravan, a houseboat, and so on.

In some cases, the receipt of multiple or duplicate questionnaires from a particular address was led by the design of the census collection strategy. For example, larger households choosing to respond on paper questionnaires had to complete and return at least one HH continuation form in addition to the primary HH questionnaire as each form was restricted to five individuals. To improve the accuracy of the census data, people were also provided with the opportunity to complete and return an individual response should they want to disclose personal information about themselves different from that returned on the primary HH questionnaire. We refer to this "individual form" as an "iForm" throughout the report. For CEs, information about the property and the residents was collected separately using a CE form and iForms. In all cases, there was no requirement to return all of the forms at the same time and no limit on the number of times a particular form could be completed.

Duplicates can also occur through respondent error. For example, in 2011, some respondents completed and returned both a paper and electronic questionnaire. Others added their details several times to the same questionnaire. Multiples and duplicates can also occur through unavoidable errors in the Census Address Frame. The Census Address Frame is a list of enumeration addresses from which we expect to receive a response. Although accurate, there are still some minor coverage and quality issues in the addressing data sources used to produce the Census Address Frame. An address recorded as being a single residential property, for example, may have recently been converted into flats, leading to the receipt of several responses from an address where only one was expected.

Importantly, multiples and duplicates can contribute to two fundamental types of error in population estimates. Leaving them in the data will inevitably lead to an overcount of properties and people. In contrast, resolving or removing them without due care and attention can lead to undercount. Failing to identify a valid response returned on the wrong questionnaire can lead to either of these errors. Both undermine the accuracy and utility of the census data, and ultimately, confidence in the census itself.

The primary role of RMR was to mitigate these risks by identifying and resolving multiples and duplicates. In most cases, after meeting the challenge of correct identification, the task was to carefully merge all of the data received from a given address into a single coherent record of the property and the occupants living there, retaining the most accurate or relevant information possible. However, the task also involved recognising when multiple responses were indicative of previously unknown addresses, and in this case, ensuring that each record was appropriately retained.

3 . Resolve multiple responses (RMR) design and development: main principles

Introduction

RMR was first implemented successfully in the 2011 Census and the general approach to the 2021 design was to build on that success. The design team included methodologists, demographers, statisticians, and other important stakeholders with a vested interest in the impact the RMR strategy would have on the data. Their role was to ensure that the development of RMR 2021 was consistent with important components of the Census 2021 collection strategy and took account of lessons learned in 2011 that identified potential improvements to the RMR process.

The final design was quality assured by two independent panels of experts: the Office for National Statistics (ONS) Census Research Advisory Group (CRAG); and an external Methodological Assurance Review Panel (MARP), both with members specialising in census methodology

As in 2011, RMR 2021 consisted of a series of modules, each focusing on a particular aspect of the problem, and building on the outcome of those before. Although based on a comprehensive review and revision of the 2011 strategy, RMR 2021 still had to overcome the same two fundamental challenges: firstly, how to identify multiple or duplicate responses, and secondly, how best to resolve them. A detailed overview of how each module tackled these challenges is provided in Section 4, Resolve multiple responses (RMR) 2021 modules: functionality and impact. Here, we provide a more general overview of the main design principles used in their development.

Identifying multiple and duplicate responses

Correctly identifying multiples and duplicates is essential. Incorrectly linking responses together (a false positive) inevitably leads to combining residences or individuals that should be counted individually in ongoing analyses. Conversely, failing to identify responses that should be linked (a false negative) leaves multiples and duplicates unresolved, leading to a residence or individual being counted more than once. Both identification errors can contribute to the potential errors in the census data rather than help resolve them.

Fortunately, some multiples and duplicates are easy to detect. For example, at a specific address, there is little ambiguity in the receipt of two household questionnaires containing exactly the same information where only one is expected. An extension questionnaire from a large household would be equally obvious. The real challenge is identifying duplicate individual responses because there are no simple indicators revealing when a person has been represented more than once in the data. To overcome this, RMR relies on "Linkage" or "Matching" methodology, designed specifically for the task. The method has to be exact enough to minimise risk of false positives, but flexible enough to avoid false negatives.

The RMR 2011 Matching methodology was based on a set of 10 match keys. Each contained a set of conditions that, if met, would indicate duplicate responses. For example, records at the address with the same name, age, and sex. To allow for differences in the spelling of a name, evaluation was supported by two text-string comparators. The Levenshtein Edit Distance method returned a score based on the similarities between letters in the name and we set a threshold for the minimum score accepted as a match. The Soundex method returned a code based on how the names sounded phonetically. Here, two names returning the same code were considered a match. Across the set of match keys, the first had the strictest conditions, but these were gradually relaxed, allowing for a small amount of acceptable uncertainty.

Development of the RMR 2021 Matching methodology identified an improved set of 17 match keys. The two strictest match keys were based on an exact match for name, date of birth, and sex. Using date of birth instead of age provided more flexibility as the rigour of the match keys was gradually relaxed. The other match keys included variations of the following conditions, either allowing some uncertainty, or ensuring issues identified in 2011 were avoided:

- date of birth either matches exactly or day or month is missing
- sex either matches exactly or is missing
- an over-30 filter to avoid matching same-sex twins incorrectly
- year-of-birth clause to prevent matching children and parents who share a forename

We retained the 2011 string-comparator, the Levenshtein Edit Distance method, for its ability to manage differences in the spelling of names:

$$\frac{Maximum\ length\ of\ first\ name - Levenshtein\ distance\ between\ first\ names}{Maximum\ length\ of\ first\ name} \geq threshold$$

However, the second string-comparator, the Jaro-Winkler Similarity method, was new for 2021, selected because of its ability to pick up a different range of errors in the name text field than the Levenshtein method:

$$Jaro - Winkler\ similarity \geq threshold$$

Together, the two string-comparators could account for a much wider range of spelling errors than those used in 2011 with improved reliability. We dropped Soundex because the review identified a risk of it generating false positives. Further detail on the RMR 2021 match keys can be found in Section 6, RMR match keys, of this methodology.

Testing and development of the RMR 2021 match keys involved detailed empirical analyses coupled with a substantial investment in clerical review. We compared performance with those used in 2011 when applied to the 2011 Census data. Overall, the research indicated that there was far less risk of false positives and false negatives using the 2021 match keys. Importantly, we were able to identify approximately 40,000 more duplicates in the 2011 Census data than those used in 2011, with a precision of over 99.99%.

Resolving multiples and duplicates: important principles

Resolving multiples and duplicates was not about arbitrarily keeping or throwing away data, but instead, about careful integration of the information available and retaining as much as possible. While there were differences in the way each RMR module tackled the problem, in general, there was a basic underlying process led by a set of business rules designed to select one response over another. We used this strategy to select the most relevant response to serve as a foundation, and then backfill any missing information in the merged response from the rest of the data available.

To define the business rules used in Census 2021, we took those used in RMR 2011 and revised them to meet many of the main principles of the Census 2021 collection design.

Electronic questionnaire (EQ) first

In 2011, information received through an EQ was prioritised over that received through a paper questionnaire (PQ). An international body of research has consistently shown that data collected electronically are more consistent and accurate than those collected on paper. Reasons include avoiding the difficult task of scanning, capturing, and coding handwritten text. As a guiding principle of the 2021 collection design, we retained this rule.

iForms first

While prioritising the information collected through individual forms was not new for 2021, more emphasis had been placed on encouraging and facilitating the use of iForms in the Census 2021 collection design than in 2011. This shift was aimed at meeting the demand for more accurate statistics on changing society norms, allowing individuals to provide personal information about themselves that they may not have disclosed on a collective household form.

Receipt date

Receipt date refers to the date we received a particular response. Although receipt date was used in 2011, for 2021 three variants were considered: receipted first; receipted last; and receipted closest to Census Day. Following the review, it was agreed that, as a main principle of the collection design was to allow people to update information by resubmitting a new response, RMR 2021 would select and retain the latest information provided, which was the response receipted last.

Most complete

In cases where there was no better reason to select one set of responses over another, it made sense to select the one with the most questions answered. Missing values are typically imputed in a later statistical process, which has its own inherent risks. Selecting the response with the least amount of missingness helps minimise the workload of later processes and risks associated with imputation.

Non-response as a valid "prefer not to say"

In 2011, calculating a completion rate was relatively straightforward, being the number of questions answered compared with the number of questions in the full census question set. However, as an extension of the demand for more accurate statistics on changing society norms, there was a legal obligation to count missing data associated with voluntary questions such as gender identity as a valid "prefer not to say" response. We factored this into all RMR modules that featured the "most complete" rule.

Communal establishment over household

The census collected information from two types of residence, communal establishments (CEs), and households (HH). In cases where both an HH and a CE questionnaire were received, we assumed that the address was a CE. Confidence was based on the fact that questionnaires were hand delivered to CEs during collection and would have been authenticated by the enumerator.

Resolving residuals

In general, within an RMR module, business rules were hierarchical, where, in the event of the first rule not being able to complete the resolution, the process would move on to the next. However, in lessons learned from 2011, it was noted that there was sometimes a handful of residual cases that remained unresolved at the end of the sequence. For RMR 2021, a statistical fallback method was implemented to resolve these residuals at random, further minimising risk of inadvertently introducing bias into the census data.

Maximising the use of all information available

Once a record had been selected as the best single source of information as a baseline using the principles outlined in this section, missing data in that record were backfilled with information from other responses received from the same person. In cases where there were several records to choose from, selection was based on the same set of principles as the baseline record.

As a final point, it is important to note that while the rules and principles outlined here seem relatively straightforward, identification and resolution of multiples and duplicates was often extremely complicated. For example, the RMR logic had to determine whether to merge or allow multiples to exist as separate households. Sometimes there appeared to be several households at an address with people evidently linked somehow by a complex matrix of relationships. On occasions, iForms could potentially be linked to more than one household. More information on this will be provided in Section 4, Resolving multiple responses (RMR) 2021 modules: functionality and impact.

New strategies for 2021

In addition to the redesign of the matching methodology and business rules, there were several opportunities to extend the scope of RMR 2021 beyond that of 2011. Here, we outline three features added to the 2021 design that substantially improved performance and subsequently, the quality of the census data.

Making use of early response data

In the 2011 Census, data capture was conducted by an external contractor. Consequently, we did not receive any data until all responses from a large, contiguous area of geography had been collected. This meant that essential RMR 2011 testing on live data did not start until 10 to 12 weeks after Census Day. For Census 2021 however, collected data were streamed daily into the Office for National Statistics (ONS) from the day the census went live, presenting a much earlier opportunity to test and quality assure the methodology. This strategy allowed time for substantial improvements to be made to the methodology and system parameters and bringing forward the delivery of deduplicated data to stakeholders and analysts.

Using alternative data sources

Making use of alternative data sources to support census processing methods was a consistent theme throughout the census programme. In 2011 it was noted that RMR could be improved with additional information about the type of residence at an address. For example, we sometimes received iForms but not a corresponding residential questionnaire. Here, information about the property from an alternative data source could be combined with the information about the individual to build a full and more accurate response.

In the RMR review, several administrative data sources were considered but, with this type of data there is always a risk that they will not be available in time. However, there was a source of information readily available in the ONS Response Management System (RMS). The RMS was designed to track responses received from each address on the Census Address Frame. In cases where we had not received a response, an enumerator was sent to the address to record basic information about the property by completing what was referred to as a "dummy form". The census data processing pipelines were extended to pull in this information and the RMR 2021 methodology was adapted to make the best use of it.

Extending the search area

In 2011, RMR focused exclusively on resolving duplicates at a discrete address. The reason for this is that extending the search to even a slightly wider geography adds a level of complexity, difficult, if not impossible, to overcome with a rule-based methodology. In the simplest case of a duplicate spanning two addresses, the difficulty starts with having to decide at which address to retain and which to delete a duplicated individual, which is made more difficult if the individual appears to be related to people in both households. The complexity of the problem grows exponentially with an increase in the search radius. Duplicates can be spread across multiple residences as well as within each residence. Leading to an inevitable complex web of relationships between all of the people involved. Fundamentally, this is the reason that the risk of duplicates having an impact on the quality of the census data, beyond those at a given address, is managed by dedicated statistical methods in later processes.

Evidence from the 2011 Census, however, suggested that there was a situation where a small extension of the RMR search could be beneficial and resolution relatively simple. Research revealed cases where all individuals in two households were duplicates. Furthermore, these households were geographically close to each other. Reasons included the situation where a house that had been converted to two flats in the past had recently been converted back to a single residence. Here, the now sole residence may have received two questionnaires, but competed them both. Alternatively, a paper questionnaire delivered to the wrong address but completed and returned in addition to a second response intended for that address, would appear in the data as a duplicate response from two different addresses.

Ultimately, the RMR 2021 search for duplicates was extended to include all residences within a postcode. In the relatively simple case of finding wholly duplicated households, we agreed that there was little or no risk in retaining only one of the households. To complete this new functionality, cases where duplicates were identified but did not cover the entire household were flagged. We used this information to help support ongoing statistical processes dealing with the more complex cases of geographically distributed duplicates.

4 . Resolve multiple responses (RMR) 2021 modules: functionality and impact

In previous sections, we have looked at the overarching role of RMR and the general principles behind its design. Here, we focus on the aims of each of the 11 RMR modules and the impact they had on the number of communal establishments (CEs), households (HHs), and individuals in the data. Before we move on, there are a few concepts worth revisiting or explaining that will make this section easier to follow.

Interchangeable terms

The terms "questionnaire" and "form" are used interchangeably to help the flow of the section, for example, "residents completed a questionnaire/form".

Counts

The counts presented here are an accurate view of the data before and after each module. However, the final output of RMR will differ from final published census outputs. There were several statistical adjustments after RMR including estimation and adjustment methods used to account for item-level and record-level non-response. An overview of these methods can be found here in our Item editing and imputation process for Census 2021, England and Wales methodology, in our Coverage estimation for Census 2021 in England and Wales methodology, and in our Coverage adjustment for Census 2021 in England and Wales methodology.

Addresses

Throughout this section we use several different terms of reference when discussing or counting places where people live. These are the main concepts:

Residential properties

Places where people live are residential properties. Communal establishments (CEs) and households (HHs) are two types of residential property. We may shorten this to just "the property" or "the residence".

Enumeration addresses

An enumeration address is an address listed on the Office for National Statistics (ONS) Address Frame used to support census collection. It is a list of addresses where we expected to receive a response from a residential property. Although extremely accurate, there were always likely to be properties missing from the list and properties on it that no longer exist.

Discovered CEs

These are residential properties we found in the data at a particular enumeration address that had been misreported as an HH with more than 30 residents.

Discovered HHs

These are multiple residential properties we found in the data at a particular enumeration address where we expected just one. This can happen, for example, where a large house had recently been converted to flats and what was 23 the High Street is now 23a, 23b, and 23c.

Residential addresses

These are the addresses of every individual residential property in the data at the end of each RMR module. For most CEs and HHs, this is the same as the enumeration address, but it also includes the address of any discovered CE or HH.

Module 1: resolution of communal establishment (CE) and household (HH) multiples

The CE questionnaire was designed to capture information about the residential property but not the individuals living there. For example, the nature of the establishment (school, prison, hospital, etc.) and who was responsible for its management (private, council, NHS, etc.). By design, we expected to receive one completed form from each enumeration address. The aim of Module 1 was to resolve cases where we received multiple CE forms, or a mixture of CE and HH forms. Residents at a CE were required to complete an iForm designed to collect information about individual people. This information would be picked up and linked back to the relevant CE in Module 5b.

For multiple CE forms, we assumed that they all related to a single residential property and were not discovered CEs at the enumeration address. On receipt of both CE and HH multiples, it was also assumed that the type of residence was a CE. This strategy was led by the fact that CE forms and iForms were hand delivered by field staff visiting the address, confirming that the residence was indeed a CE and not an HH.

Overall, we received at least one completed CE questionnaire from 35,723 different properties with a unique residential address. However, from around 6,000 of those, we also received an additional 240 CE and 19,874 HH questionnaires likely either to be duplicates or respondent error. Table 1 shows how the multiple responses were distributed.

Table 1: Number of addresses with multiple CE and HH forms

Type of multiple response	n	%
Two or more CEs and no HHs	83	0.23
Two or more CEs and at least one HH	45	0.13
One CE and at least one HH	5,864	16.42
Total	5,992	16.77

Source: Office for National Statistics

Notes

1. N = 35,723

In general, the receipt of multiple CE forms was rare, with only 240 cases spread over 128 different residential addresses. Given the questionnaire delivery method, we expected the receipt of duplicate CE forms to be low like this. In contrast, we did not expect to receive any HH forms from a CE address but ended up with almost 20,000. Further investigation revealed that when handing out iForms to individuals at student halls of residence, there were instances where field staff inadvertently gave out HH forms instead. Analyses confirmed that in some cases, over 100 HH forms were received from the same CE. In total, information for 58,278 individuals was captured on these HH forms. Module 1 was designed to convert this information into iForms to be picked up correctly in Module 5. Overall, Module 1 did not have any impact on the number of CEs in the data. However, we return to the issue of discovered CEs in Module 11.

Modules 2 and 3: resolution of multiple HH responses and discovered HHs

In a way similar to the CE questionnaire, the HH questionnaire was designed to capture information about the residential property. For example, whether it was a house or a flat; the number of rooms it had; and so on. However, unlike the CE form it was also designed to capture information about each individual living there. Again, by design, only one completed HH form was expected from a given enumeration address. The aim of Modules 2 and 3 was to resolve only the information about the residential property. The possibility of duplicate individuals represented on multiple questionnaires received from the same address would be tackled in Modules 7 and 8.

Unlike CE forms, multiple HH forms could be resolved in two ways. As outlined in Section 2, The role of resolve multiple responses (RMR), there were several reasons why a respondent might complete and return more than one HH questionnaire. In all cases, it was extremely likely that all the information provided belonged to one HH. This meant that to avoid an overcount of HHs in the data, the information should be merged. In contrast, because of the potential inaccuracies in the Census Address Frame mentioned at the beginning of this section, multiple HH forms might be indicative of there being more than one residential address at the enumeration address. Here, merging multiple responses would not be the correct thing to do as this would lead to undercount. Consequently, we defined a sequence of rules to identify cases where multiple HH forms could be merged into one with a high level of confidence. In situations where these conditions could not be met, multiple responses were considered to be discovered HHs and these were retained as such in the data.

Merge rule 1: the same questionnaire identifier

In cases where a respondent made a point of requesting a second questionnaire in a different format, the new form would be given the same questionnaire identifier as the initial response to help ensure they could be linked together in RMR. For example, where they first requested an electronic questionnaire, and then later asked for a paper version, or the other way around.

Merge rule 2: duplicate individuals

Where questionnaire identifiers differed, we used a matching exercise to identify duplicate individuals captured in each set of HH forms received from a given address. The details of every person were compared with those of every other person using the match keys described in Section 3, Resolve multiple responses (RMR) design and development: main principles. Questionnaires with at least one duplicate person in common were considered to be from the same HH.

Merge rule 3: under 16s only

Following the matching exercise, responses were checked to see if any of the questionnaires comprised only individuals under the age of 16 years. An HH without any adults was deemed very unlikely, and any forms like this were merged with another HH response containing an adult with the closest matching surname.

Merge rule 4: related individuals

Multiple responses from a given address containing individuals who shared the exact surname were merged on condition that information about the property itself, such as number of bedrooms, was also the same. Here, we considered highly likely that the additional HH forms were completed instead of a continuation form. Please find more information about continuation forms in Module 5.

Merge rule 5: English and Welsh responses

Brought forward from the 2011 Census, we assumed that HH forms received on both the English and Welsh versions of the questionnaire were from the same HH. Research indicated that the most likely reason for this combination was that people receiving an English form but wanting to reply in Welsh might ask for a second form but written in their preferred language.

Merge rule 6: empty properties

HH forms with no people were assumed to be duplicates of an occupied HH response at the same address. These records were merged as long as there were similarities in the information on the forms, such as number of bedrooms. Cases where all of the properties were empty were merged in a similar way.

Overall, we received just over 25 million HH questionnaires. However, Modules 2 and 3 discovered that just under 560,000 of those were duplicates and these were merged accordingly. In contrast, we also identified around 97,000 discovered households with a residential address different than the original enumeration address. These were retained in the data. Table 2 provides a detailed summary of how HH forms were resolved and the impact that had on the total number of HHs in the data.

Table 2: Resolution of multiple HHs forms and overall impact on HH count

	n
Number of HH questionnaires in the data at the start of Modules 2&3	25,354,898
HH forms identified as being from the same HH and merged	558,322
HH forms identified as discovered HHs and retained	96,910
Final number of HHs in the data after Modules 2&3	24,796,576

Source: Office for National Statistics

Ongoing quality assurance and review from dedicated teams of researchers confirmed and validated the retention of discovered HHs and their corresponding residential addresses. Merging the duplicate forms and retaining the discovered HHs led to there being just under 25 million HHs in the data at the end of this process. By the end of Modules 2 and 3, all multiple CE and HH forms had been resolved.

Module 4: resolution of dummy forms

Dummy forms were electronic questionnaires completed by census field staff. Every questionnaire provided to a potential respondent had a reference number that related to the enumeration address on the Census Address Frame. This meant that we could keep track of those addresses where we had not received a completed response. Although time and resources made it impossible to visit every address yet to return a form, field staff were sent to as many as possible. If unable to contact someone living there, they would collect basic information about the property. For example, whether it looked like a CE or an HH, the type of accommodation it appeared to be, and an estimate of the number of people likely to be living there.

The primary role of Module 4 was to add the information collected on dummy forms to the census data to support the statistical methods used after RMR to account for item- and record-level non-response mentioned at the beginning of this section and counter the problem of potential undercount. The information provided on dummy forms helped to remove some of the burden placed on these methods, improving their accuracy, and ultimately, the final census outputs. As a discussion about these methods is beyond the scope of this report, we refer the interested reader to the methodology links given under "Counts", at the beginning of this section.

For some addresses we collected a dummy form in addition to a completed HH or CE questionnaire. This happened simply because the respondent filled in and returned the questionnaire after the property had been visited. In these cases, we ignored the dummy form because the information provided by the respondent was likely to be more accurate. Overall, we collected 2,679,041 individual dummy forms although many of these were duplicates created by several return visits to the property in an effort to get a response. Multiple dummy forms from the same address were merged into one, again using the principles outlined in Section 2, The role of resolve multiple responses (RMR).

Table 3 shows how the integration of dummy forms had an impact of the number of HHs and CEs in the data.

Table 3: Integration of dummy forms and overall impact on CE and HH count

	n HHs	n CEs
Properties at the end of Module 3	24,796,576	35,723
Dummy forms added in Module 4	1,537,794	4,016
Properties at the end of Module 4	26,334,370	39,739

Source: Office for National Statistics

Altogether, Module 4 added just over 4,000 CEs and 1.5 million HHs to those retained at the end of Module 3. This represented an 11.24% increase in the number of CEs in the data and 6.20% for HHs. At the end of Module 4 there was a unique and discrete set of information for just over 39,700 CE and 26.3 million HH residential addresses.

Module 5a: resolution of household continuation forms

By design, continuation forms were associated with respondents choosing to complete an HH paper questionnaire (PQ). Unlike the online electronic HH questionnaire, the PQ only had space for information about five individuals and respondents from larger HHs would have to request a continuation form. Continuation forms were also limited to five individuals, meaning that we might receive several from a particularly large HH. The primary aim of this module was to link people on these forms to one of the HHs retained in the data from previous modules.

Largely, Module 5a was relatively straightforward. In cases where the form had the same residential address as a unique HH retained or added by previous modules, people on the form could simply be assigned to that HH. In cases where a continuation form had the same address as a CE, we assumed that the respondent most likely misunderstood their response options and incorrectly completed this form instead of, or as well as, the iForm required. Here, the principles outlined in Module 1 were brought forward and the respondent's data were converted into an iForm to be linked to the CE in Module 5b.

A third and more difficult case occurred where Modules 2 and 3 had determined that there were discovered households at an enumeration address, each with a different residential address, but the continuation form contained only the enumeration address. Here, the challenge was to link the form, and the people on it, to the residential address where they most likely lived. To achieve this, the matching strategy outlined in Section 3, Resolve multiple responses (RMR) design and development: main principles, was applied again, to identify commonalities between people in each of the discovered HHs and on each continuation form.

The resolution strategy took account of duplicates where an individual appeared to be represented more than once across the set of forms; people likely to be related because they have the same or a similar sounding surname; and people who had nothing in common at all. Continuation forms were ultimately appended to the discovered HH where the number of matching individuals and the overall quality of any matches was the highest.

Overall, we received a total of 7,815 HH continuation forms containing information about 15,251 people. It is important to note that as with previous modules, we made no effort to resolve any duplicate individuals identified at this point but we looked at this again in Modules 7 and 8. Table 4 shows how the continuation forms were distributed across the HHs and CEs retained in the data by previous modules.

Table 4: Resolution of household continuation forms

	Forms		Persons	
	n	%	n	%
Cont. Forms assigned to a HH	7,262	92.92	14,216	93.21
Cont. Forms assigned to a CE	24	0.31	89	0.58
Cont. Forms unassigned	529	6.77	946	6.2
Total	7,815	100	15,251	100

Source: Office for National Statistics

With around 93% of the forms being linked to an HH residential address, consistent with their intended purpose, it was clear that most people requesting and completing a continuation form did so correctly, and as expected. There was evidence that a few people living in a CE misunderstood the design but with only 24 cases, the number of forms that linked to a CE address, and the number of people converted to iForms, was extremely low. Overall, where continuation forms could be linked to an HH or CE, we were extremely confident that the information collected on those forms had been allocated correctly.

The 529 forms and 946 individuals that could not be assigned to an HH or CE each came with an address different to any of those recorded on HHs or CEs retained in the data. Here, we assumed that it was highly likely that there was an intention to provide a corresponding HH form, but for some reason we did not receive it. As continuation forms were primarily associated with paper questionnaires, it was quite possible that the respondent simply forgot to post it. Not wanting to lose this information, we return to these forms again in Module 6.

Module 5b: resolution of iForms

As outlined in Section 2, The role of resolve multiple responses (RMR), iForms played an important role in the Census 2021 collection strategy. The general aim was to give all individuals the opportunity to provide or update personal information about themselves as confidentially as possible, even if they had been included on another questionnaire. As discussed in Module 1, iForms were also the preferred method for capturing information about individuals living in a CE. The overarching role of Module 5b was to assign iForms to HHs or CEs retained in the data by previous modules.

The primary task here was similar to that for continuation forms where each iForm was simply linked to a CE or HH with the same residential address. In the more difficult case where we had iForms with an enumeration address that matched that of a set of discovered HHs with different residential addresses, we used the same matching strategy applied to continuation forms to determine in which discovered HH they most likely lived. Again, the person of each iForm was compared with all of those in the discovered HHs and assigned to the one containing an individual, or individuals, with the most features in common. Here, however, there was an additional risk.

With multiple iForms there was a possibility that some of them were duplicates associated with the same individual. Subtle differences in the information provided on duplicate forms could, in principle, lead them to being linked to different HHs. This would make it impossible for them to be identified as duplicates within a residential property as planned in Modules 7 and 8 and mean that they might still contribute to unwanted overcount in census outputs.

To overcome this problem we implemented a strategy to identify and resolve duplicate iForms at each address prior to assigning them to a residential address. Individuals on multiple iForms captured at the same address were compared with each other using the match keys used in other modules and described in Section 3,Resolve multiple responses (RMR) design and development: main principles. The iForms meeting the criteria to be considered duplicates were resolved into one coherent response. The business rule hierarchy used firstly, to identify the foundation record and then secondly, any records needed for backfilling missing information was as follows:

Receipt date

Responses with the latest receipted date took top priority.

If all duplicate responses were receipted on the same date, we looked to the level of completion.

Level of completion

The most completed responses took priority.

If all duplicate responses were completed to the same level, we looked to the form type.

Form type

Respondents had the option to fill out either an electronic questionnaire or paper questionnaire. Responses captured on electronic forms took priority over paper forms.

If all duplicate responses were captured on the same form type, we selected one at random.

Table 5 shows how many iForms we had in the data and the impact of the deduplication process.

Table 5: The impact of deduplicating iForms

	n	%
iForms added by Module 1	58,278	5.25
iForms added by Module 5a	89	0.01
iForms returned by respondents	1,051,765	94.74
Total Number of iForms in the data	1,110,132	100
iForms deduplicated and merged	12,516	1.13
Final Number of iForms retained	1,097,616	98.87

Source: Office for National Statistics

At the beginning of Module 5b, we had just over 1.1 million iForms in the data. Just under 95% of those were authentic returns with the rest being information about individuals we recovered successfully from resolving CEs in Module 1 and continuation forms in Module 5a. The deduplication process identified and removed around 12,500 duplicates. At just over 1%, this was a relatively small amount, but big enough to contribute to the problem of overcount in census outputs if not removed, supporting the decision to remove them before linking them to their respective residential properties.

There were just under 1.1 million iForms in the data at the end of deduplication. Table 6 shows how they were distributed.

Table 6: iForms assigned to residential addresses in Module 5b

	n	%
iForms assigned to a CE	832,882	75.88
iForms assigned to a HH	225,732	20.57
iForms unassigned	39,002	3.55
Total	1,097,616	100

Source: Office for National Statistics

Overall, approximately 76% of all iForms were linked to a CE. As this was the primary method for collecting information about individuals living in this type of residence, this made perfectly good sense. With just over a quarter of a million iForms being assigned to HHs, it seemed that there were quite a few individuals keen to provide or update information about themselves that may have differed from that on the main HH questionnaire. This clearly justified encouraging people to use iForms in the Census 2021 collection design to help improve the accuracy of information collected, as outlined in Section 2, The role of resolve multiple responses (RMR).

In a way similar to continuation forms, there were just over 39,000 iForms we could not assign because they had an address different to the enumeration or residential address of any the HHs or CEs retained in the data. Again, we assumed that an intention to provide a corresponding HH or CE form was very likely, but for some reason we did not receive it. Along with the unassigned continuation forms from Module 5a, we return to the unassigned iForms in Module 6.

Module 6: resolution of residual responses

At the end of Modules 5a and 5b, there were a number of continuation forms and iForms that could not be assigned to a CE or an HH because they had an address that differed from any of those retained in the data by previous modules. These questionnaires and the people on them were referred to as residual responses. As mentioned previously, in all cases, we assumed that there was an intention to provide a respective CE or HH form, hence, the aim of Module 6 was to create appropriate HH or CE responses at each residential address provided on these forms and link the information about the residual individuals to them.

It is important to note here that although we were able to distinguish between a CE and an HH from the information on the residual questionnaires, we did not have any of the other information that would have been collected on the standard CE and HH forms such as the type of accommodation, tenure, number of rooms, and so on.

Here, however, this strategy had the same objective as adding dummy forms to the data as we did in Module 4. The information would serve to reduce the burden and substantially improve the accuracy of ongoing statistical estimation and adjustment processes designed to account for item- and record-level non-response. Again, we refer the interested reader to the links provided under "Counts" at the beginning of this section.

To recap, at the end of Modules 5a and 5b, we had 529 residual continuation forms containing 946 individuals and 39,002 individual iForms. Table 7 shows how Module 6 resolved them by creating and adding new CE and HH forms to the data and assigning individuals to the appropriate residential address.

Table 7: The distribution of Residual forms

	New residences	Individuals assigned
CE	6,021	14,458
HH	23,223	25,490
Total	29,244	39,948

Source: Office for National Statistics

Overall, Module 6 recovered just over 29,000 residential properties that would have otherwise been lost. This included approximately 6,000 CEs containing just under 14,500 individuals, and around 23,200 HHs containing almost 25,500 individuals.

Although we make some final adjustments to the residential property counts in Modules 9,10, and 11, the residual forms represented the last of the questionnaires captured during the census collection period containing information about CEs and HHs not yet linked to the data retained by previous modules. To mark that, Table 8 shows the number of properties in the data prior to, and after, Module 6.

Table 8: The number of residential properties in data at the end of Module 6

	Pre Mod. 6	Post Mod. 6	% increase
CE	39,739	45,760	15.15
HH	26,334,370	26,357,593	0.09
Total	26,374,109	26,403,353	0.11

Source: Office for National Statistics

Overall, Module 6 captured just over 0.11% of the total number of residential properties, but perhaps most importantly, 15.5% of CEs. Without this adjustment it was quite clear that there would have been a substantial undercount of the number of CEs, potentially leading to errors in any census outputs based on this type of property. By the end of Module 6, there were around 45,800 CEs and 26.4 million HHs retained in the data.

Module 7 and 8: identification and resolution of duplicate persons within the same residence

Although all of the modules discussed so far have played an important role in minimising the risk of undercount and overcount in the census data, and ultimately, the accuracy of any outputs and analyses, duplicate individuals were likely to have a big impact on overcount and population estimates if left unresolved. As mentioned in other modules, up until this point, apart from the resolution of iForms, we have not looked for duplicate individuals within either a CE or an HH, even when adding additional records to these properties. The aim of Modules 7 and 8 was to remedy this.

As in other modules, the primary method used to identify duplicates within a residential address was the matching method described in Section 3, Resolve multiple responses (RMR) design and development: main principles. As before, the information collected about each individual at a given address was compared with that of every other individual. Where two or more records met the thresholds set to be considered a match, the group of responses were considered to be duplicates and the information considered to belong to the same person. It is important to note that this module was designed to identify any number of duplicate individuals within a residence. For example, an HH starting with information about six individuals could, in principle, be resolved to a three-person HH based on three forms belonging to person 1, two forms belonging to person 2 and the final form belonging to person 3.

The business rule hierarchy we used here that allowed us to identify which record to retain as the foundation and which to use in backfilling missing information was as follows:

iForms

iForms took priority over all other responses.

If all duplicated responses were iForms, we looked to the level of completion.

Level of completion

The most completed responses took priority.

If all duplicate responses were completed to the same level, we looked to the form type.

Form type

Respondents had the option to fill out either an electronic questionnaire or paper questionnaire. Responses captured on electronic forms took priority over paper forms.

If all duplicate responses were the same form type, we looked to the receipt date.

Receipted date

Responses with the latest receipted date took priority.

If all duplicate responses were receipted on the same date, we selected one at random.

Table 9 shows how many duplicate individuals were identified and resolved into a single, coherent response in Modules 7 and 8.

Table 9: The number of duplicate individuals identified and resolved

	Resolved duplicates
CE	7,417
HH	741,770
Total	749,187

Source: Office for National Statistics

At just over 7,400 records, there were relatively few duplicate individuals linked to CEs. In contrast, we identified just under three-quarters of a million duplicates at a residential HH address. This number would have had a substantial impact on the accuracy of any population outputs based on the census data had they not been identified and resolved.

In a way similar to the previous module and residential property counts, although we make some final adjustments to the numbers of individuals in the data in Modules 9 and 10, the deduplication of individuals within residence represented the last adjustment based directly on information captured on distinct questionnaires related to the collection design. Table 10 shows the number of individuals in the data prior to, and after this module.

Table 10: The number of individuals in the data at the end of Module 7 and 8

	Pre Mod. 7	Post Mod. 8	% decrease
CE	847,340	839,923	0.88
HH	58,564,315	57,822,545	1.27
Total	59,411,655	58,662,468	1.26

Source: Office for National Statistics

Overall, approximately 1.3% of the total number of individuals in the data prior to Modules 7 and 8 were identified as being duplicates and resolved into single record. By the end of Modules 7 and 8, there were approximately 58.7 million individuals in the data, with around 840,000 living in CEs and just over 57.8 million living in HHs.

Modules 9 and 10: resolution of duplicate residences within the same postcode

Modules 9 and 10 were a final attempt to identify and reduce potential duplicates in the data. The primary aim was to look for entirely duplicated residential properties within a postcode. Recognised as an issue in 2011, this was a new module for RMR 2021 made possible by a substantial increase in processing power. More detail about the background and design can be found in Section 3, Resolve multiple responses (RMR) design and development: main principles. A duplicate residence was considered to be two or more properties containing the same people. This was something identified in 2011 that might occur when the Census Address Register contained two different enumeration addresses, but, for some reason, such as a recent local building programme, both corresponded to just one residential address. Here, we might receive two questionnaires that appear to come from two different addresses but in fact they only belong to one.

To identify a duplicate residence, the information about the CE or HH and all of the individuals living in a particular property were compared with that of other properties within the postcode. The matching strategy outlined in Section 3, Resolve multiple responses (RMR) design and development: main principles, and employed in other modules, was used again to compare the characteristics of the individuals. We repeated this process, comparing every property within the postcode, and for all postcodes in England and Wales.

Compared with other modules, the business rules for resolving wholly duplicated residences had to be much simpler. Individual responses at this stage could be the result any of the processes applied in previous modules and taking this into account would have been far too complex. Consequently, we simply retained the residence containing the most information. Table 11 shows how many duplicate residences and individuals were identified in the data.

Table 11: Duplicate residences identified and resolved within postcode

Record type	Residences	Individuals
CE	9	86
HH	20,219	38,670
Total	20,228	38,756

Source: Office for National Statistics

Given that prior to this module there were around 26.5 million properties and 58.6 million individuals in the data, the number of duplicate residences and individuals living within them represented an overall change of less than 0.01%. However, a count of just over 20,000 duplicate residences and almost 39,000 individuals would still have a substantial impact on the accuracy of census outputs. Moreover, resolving these duplicates here is a far better strategy than leaving them in the data and trying to manage them statistically in later parts of the process, as happened in 2011. Overall, the decision to widen the search area for duplicate residences within the same postcode proved extremely worthwhile.

Module 11: identification and resolution of discovered CEs

Module 11 was the last in RMR, designed to identify CEs in the data that may have been misclassified as an HH. This functionality was new to RMR 2021, but the problem was recognised and a solution implemented in the 2011 Census as a stand-alone statistical adjustment method after RMR.

This possibility of misclassified CEs can arise as a result of the enumeration address not being listed as CEs in the Census Address Frame in addition to the CE manager completing and returning an HH form rather that the CE form and associated iForms as outlined in Module 1. Unlike other CEs, these properties would not have been visited by an enumerator making the respondent error difficult to detect.

Research in 2011 concluded that these hidden CEs could be discovered with a high level of confidence by looking for HHs that had more than 30 residents. It also concluded that other HHs at the enumeration address, despite having a different residential address, were more than likely linked to the CE than being a discovered HH retained in Modules 2 and 3. For example, a discovered CE might be student accommodation consisting of a number of individual flats.

To resolve this problem, Module 11 was designed to first identify HHs with more than 30 residents. Once discovered, the information relating to that property, and any other residential property at the same enumeration address was converted into a single, unique CE, ensuring that all of the residents were also assigned to the CE correctly. Table 12 shows how many CEs we discovered in the data and the number of misclassified HHs we collapsed back into this correct category.

Table 12: Discovered CEs and misclassified HHs

Record type	Residences	Individuals
CE	188	18,822
HH	-1,175	-18,822
Total	-987	0

Source: Office for National Statistics

Overall, we discovered 188 CEs across England and Wales at an enumeration address originally classified as being an HH with more than 30 residents. We also found a total of 1,175 HHs and almost 19,000 individuals related to those addresses. Although these counts seem low, they can make a considerable difference to how local communities are represented in the data. For instance, at one specific address, there were 305 individual HH responses that were merged to form a CE. Had we not implemented Module 11, this would have been immediately evident in census outputs and analysis to anyone with local knowledge of the CE in question.

5 . Evaluation

As we have provided a detailed evaluation of all components in the resolve multiple responses (RMR) process in previous sections, here, we provide a final high-level evaluation of the performance of the end-to-end RMR strategy with an overview of how the process met the main aims and objectives set out at the beginning of the report.

As mentioned in previous sections, the census collection strategy was designed to maximise the accuracy of the census data by providing as many opportunities for people to respond as possible. This, and other collection design decisions, led to there being five fundamental types of census questionnaire. Notably, information relating to a particular communal establishment (CE) or household (HH) could be collected across several forms and there were no constraints on the number of times someone could respond. Table 13 shows the structure of the raw data at the beginning of the RMR process and how many responses we had on each type of questionnaire.

Table 13: Questionnaires in the data at the beginning of RMR

	Residences	Individuals
CE Forms	35,963	0
HH Forms	25,374,772	58,357,155
Dummy Forms	2,679,041	0
Continuation Forms	0	15,251
iForms	0	1,051,765
Total	28,089,776	59,424,171

Source: Office for National Statistics

The pre-RMR data contained just over 28 million records with information relating to a residential property and 59.4 million records containing information about individual people. At this stage, as expected, CEs had no residents, larger HHs who responded on paper questionnaires were limited to five individuals, and over 1 million individuals had not yet been linked to an HH or CE. The effort put into improving response rates by field staff was also very evident in the 2.7 million dummy forms collected. Importantly, as discussed throughout the report, the data also contained duplicates and other response anomalies likely to contain important information about residential properties and individuals not captured exactly as intended.

Implemented in Modules 5a and 5b, the simplest task for RMR was to assign the information collected about individuals on continuation forms and iForms to the appropriate CE or HH. Table 14 shows how many individuals were assigned to a residential property through this planned linkage strategy.

Table 14: Individuals assigned to a residential property

Module		Individuals
5a	Resolve Continuation Forms	14,216
5b	Resolve iForms	1,058,614
	Total	1,072,830

Source: Office for National Statistics

Overall, RMR successfully assigned just over 1.1 million individuals to a residential property. There were some complications within this process, such as the need to generate additional iForms in Module 1 and resolution of residual iForms and continuation forms with no corresponding CE or HH questionnaire in Module 6. However, this was largely about matching the address recorded on either of these forms to that of a residential property, as planned in the design.

As discussed in detail throughout the report, the most important role of RMR was to identify and resolve errors and anomalies in the data to improve both quality and utility. Here, we recall that there were two fundamental issues addressed by the RMR process likely to have a detrimental impact on census outputs and analysis, namely, overcount and undercount. The main contribution to the risk of overcount comes from multiple or duplicate responses relating to the same property or individual. The main contribution to the risk of undercount, in this context, comes from not identifying, recovering, and retaining discovered CEs and HHs, or information about residential properties and individuals responding in ways that are not consistent with the collection design. Table 15 provides an overview of the duplicates and recovered responses identified and resolved by each of the RMR modules.

Table 15: Duplicates and Recovered responses identified and resolved by RMR

Module		**Resolved Duplicates: Overcount**		**Recovered Responses: Undercount**	
		Properties	**Individuals**	**Properties**	**Individuals**
1	Resolve CEs 1	20,114			58,278
2&3	Resolve HHs	558,322		96,910	
4	Resolve Dummies	1,137,231		1,541,810	
5b	Resolve iForms		12,516		
6	Resolve Residuals			29,244	39,948
7&8	Resolve individuals		749,187		
9&10	Resove residences	20,228	38,756		
11	Resolve CEs 2	1,175		188	
	Total	1,737,070	800,459	1,668,152	98,226

Source: Office for National Statistics

From Table 15, the importance of RMR and the role it played in improving the accuracy of the census data is quite evident. The identification of duplicates using the new matching methodology outlined in Section 2, The role of resolve multiple properties (RMR), and other strategies applied in many of the RMR modules, identified and resolved just over 1.7 million multiple or duplicate residential properties and just over 800,000 duplicate individuals. These records would have contributed to a substantial amount of overcount in census analyses and outputs if left untreated. Conversely, RMR was able to identify and recover almost 1.7 million properties and just over 98,000 individuals that otherwise would not have been represented in the data. This would have led to a significant amount of undercount had we not invested a considerable amount of effort identifying ways of recognising and collecting valuable information reported beyond the basic collection design. Table 16 shows the structure of the data at the end of RMR.

Table 16: The final structure of the census data after RMR

	Residences	**Individuals**
CE	45,939	858,659
HH	26,336,199	57,765,053
Total	26,382,138	58,623,712

Source: Office for National Statistics

In general, RMR was an extremely complex process that had a considerable amount of time and effort invested in its design and development. However, we are confident that RMR met all of its primary aims and was worth the effort. The multiples, duplicates, and the types of errors we identified and resolved here are, in many ways, a necessary consequence of a collection strategy designed to optimise and maximise response. However, coupled with a well-designed RMR process, built in conjunction with other critical statistical processes that follow on from RMR, we were able to minimise the impact of these issues and errors and provide consumers of the census data and stakeholders with a clean and accurate database from which to build high-quality population estimates. Importantly, RMR helped improve the utility of the data at much lower levels of aggregation where duplicates are likely to be far more salient, potentially undermining trust in the census. Overall, we believe that RMR made a crucial contribution toward the overarching aim of the census to provide detailed, high-quality information about the population in England and Wales that will ultimately be used for the public good.

6 . Resolve multiple responses (RMR) match keys

Match key 1 rules

- Address ID or household ID matches exactly
- Full name matches exactly, including middle name
- Date of birth matches exactly
- Sex matches exactly

Match key 2 rules

- Address ID/ household ID matches exactly
- First name matches exactly
- Surname matches exactly
- Date of birth matches exactly
- Sex matches exactly

Match key 3 rules

- Address ID/ household ID matches exactly
- First name matches exactly
- Surname matches exactly
- Year of birth matches exactly
- Month of birth matches exactly
- Error in day
- Sex matches exactly

Match key 4 rules

- Address ID/ household ID matches exactly
- First name matches exactly
- Surname matches exactly
- Year of birth matches exactly
- Day of birth matches exactly
- Error in month
- Sex matches exactly

Match key 5 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- Full name has a Levenshtein distance less than 3, but excludes names less than 3 characters long
- Date of birth matches exactly
- Sex matches exactly

Match key 6 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name has a Jaro-Winkler similarity distance greater than 0.7 (may vary by module)
- Surname has a Jaro-Winkler similarity distance greater than 0.7 (may vary by module)
- Date of birth matches exactly
- Sex matches exactly

Match key 7 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name has a Levenshtein edit distance greater than 0.6 (may vary by module)
- Surname has a Levenshtein edit distance greater than 0.6 (may vary by module)
- Date of birth matches exactly
- Sex matches exactly

Match key 8 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name and surname transposed
- First name/ surname Jaro-Winkler similarity distance greater than 0.7 (may vary by module)
- Surname/ first name Jaro-Winkler similarity distance greater than 0.7 (may vary by module)
- Date of birth matches exactly
- Sex matches exactly

Match key 9 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name and surname transposed
- First name/ surname Levenshtein edit distance greater than 0.6 (may vary by module)
- Surname/ first name Levenshtein edit distance greater than 0.6 (may vary by module)
- Date of birth matches exactly
- Sex matches exactly

Match key 10 rules

- Address ID/ household ID matches exactly
- Allows under 30 years of age
- First name has a Jaro-Winkler similarity distance greater than 0.9 (may vary by module)
- Surname has a Jaro-Winkler similarity distance greater than 0.9 (may vary by module)
- Date of birth matches exactly
- Sex matches exactly

Match key 11 rules

- Address ID/ household ID matches exactly
- Allows under 30 years of age
- First name has a Levenshtein edit distance greater than 0.9 (may vary by module)
- Surname has a Levenshtein edit distance greater than 0.9 (may vary by module)
- Date of birth matches exactly
- Sex matches exactly

Match key 12 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name has a Jaro-Winkler similarity distance greater than 0.85 (may vary by module)
- Surname has a Jaro-Winkler similarity distance greater than 0.85 (may vary by module)
- Date of birth matches exactly
- Sex is missing or matches exactly

Match key 13 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name Levenshtein edit distance greater than 0.85 (may vary by module)
- Surname Levenshtein edit distance greater than 0.85 (may vary by module)
- Date of birth matches exactly
- Sex is missing or matches exactly

Match key 14 rules

- Address ID/ household ID matches exactly
- Over 30 years old
- First name Levenshtein edit distance greater than 0.85 (may vary by module)
- Surname Levenshtein edit distance greater than 0.85 (may vary by module)
- Date of birth has a Levenshtein distance of less than 2 (may vary by module)
- Sex matches exactly

Match key 15 rules

- Address ID/ household ID matches exactly
- First name matches exactly
- Surname matches exactly
- Year of birth matches exactly
- Day and month of birth transposed
- Day/ month matches exactly
- Sex matches exactly

Match key 16 rules

- Address ID/ household ID matches exactly
- First name Levenshtein edit distance greater than 0.9 (may vary by module)
- Surname Levenshtein edit distance greater than 0.9 (may vary by module)
- Year of birth matches exactly
- Month of birth matches exactly
- Error in day
- Sex matches exactly

Match key 17 rules

- Address ID/ household ID matches exactly
- First name Levenshtein edit distance greater than 0.9 (may vary by module)
- Surname Levenshtein edit distance greater than 0.9 (may vary by module)
- Year of birth matches exactly
- Day of birth matches exactly
- Error in month
- Sex matches exactly

7 . Related links

Item editing and imputation process for Census 2021, England and Wales - Office for National Statistics (ons.gov.uk)
Methodology | Last revised 8 November 2022
The methods for resolving item non-response and item inconsistencies in Census 2021 data, including deterministic editing, nearest neighbour donor imputation, and manual imputation.

Coverage estimation for Census 2021 in England and Wales - Office for National Statistics (ons.gov.uk)
Methodology | Last revised 9 November 2022
Methodology for coverage estimation of Census 2021 in England and Wales.

Coverage adjustment for Census 2021 in England and Wales - Office for National Statistics (ons.gov.uk)
Methodology | Last revised 19 December 2022
Methodology for the coverage adjustment of Census 2021 in England and Wales.

Evaluation of addressing quality: Census 2021 - Office for National Statistics (ons.gov.uk)
Methodology | Last revised 9 January 2023
Summary of the work conducted to build the Census 2021 address frame and an evaluation of the frame's quality.

8 . Cite this methodology

Office for National Statistics (ONS), released 22 June 2023, ONS website, methodology, Improving the accuracy of the Census 2021 data by resolving multiple responses (RMR)