Maximising the quality of Census 2021 population estimates

How we maximised the quality of Census 2021 population estimates during the processing and quality assurance of the final statistics.

Contact:
Census customer services
census.customerservices@ons.gov.uk
+44 1329 444972

Release date:
9 December 2022

Next release:
To be announced

Table of contents

1 . Main points

- Census 2021 exceeded quality targets as detailed in Sections 1.7 and 1.8 of our census White Paper, Help Shape Our Future.
- We had target person response rates of 94% overall and at least 80% in every local authority (LA) and we achieved 97% overall and over 88% in every LA.
- Our planned flexible approach to collection and well-tested response strategy enabled us to respond to changing circumstances, such as the coronavirus (COVID-19) pandemic.
- Our data quality control strategy functioned to detect and then fix any data issues throughout the collection and processing stages.
- Our coverage estimation approach built on what we had done in previous censuses; it allowed us to estimate and adjust for non-response and provide a fully imputed and consistent dataset based on those estimates.
- Our data validation strategy helped to detect and fix any implausible estimates by analysing coherence between census, and other data and statistics.
- For the first time, we asked local government organisations to help us assess the data before publication, with over half of these participating in this process.

2 . Preparing to meet Census 2021 data needs

Census 2021 took place in England and Wales around Census Day on 21 March 2021. Run by the Office for National Statistics (ONS), it was the first digital-by-default census in the UK.

The census provides a unique snapshot of our society every 10 years, giving us the most accurate estimate of all the people and households in England and Wales. Information from the census helps the government and local authorities (LAs) to plan and fund local services, such as education, healthcare and transport.

Criteria for the quality of census statistics were set out in the December 2018 White Paper, Help Shape Our Future (PDF, 967KB), and in our Design for Census 2021 article, published in October 2020. These included response rate targets of 94% for England and Wales as a whole and no less than 80% for any LA, and challenging targets to minimise variability. With a final person response rate of 97% and at least 88% in all LAs, these targets were exceeded.

Census 2021 took place during the coronavirus (COVID-19) pandemic. In October 2020, we also published our Operational Planning Response to the coronavirus pandemic article. This detailed how we planned to deliver the census to meet these success requirements while considering the potential impact of the pandemic.

This article describes how we built on the success of the collection phase to maximise the quality of the final census estimates through our statistical processing and quality assurance. Alongside this we are also publishing our Quality and methodology information (QMI) for Census 2021 report, which discusses in more detail the quality measures of census data.

No census is perfect, and our approach has always been planned to include methods to measure and account for non-response and uncertainties in collection and processing.

3 . Collecting the data

The design for Census 2021 was optimised to meet our targets. These targets were for overall response levels and to minimise the level of variability across England and Wales.

Some important features of the design to help meet these targets included:

- a “wave of contact” approach, which had been rehearsed in 2019 to phase initial contact and follow-up activity, as detailed in our 2019 Collection rehearsal evaluation report for Census 2021, England and Wales
- a “Hard to Count” strategy, which prioritised greater resource in areas identified as being likely to require greater follow-up activity
- identification of “Target Action Groups” where the standard approach needed to be adapted to enable households and individuals to complete
- detailed management information updated daily to track return rate levels against expectations and to adapt follow-up activity as required
- ensuring everyone had the means to respond, enabled through design and extensive testing of the online and paper questionnaires and provision of extra telephone capture capacity
- sending paper questionnaires as initial contact to around 10% of areas where households might be willing to respond but less able to respond online

We achieved a 97% overall response rate, exceeding our target of 94%.

Once the data were collected, they went through a series of processing steps before being ready to publish. These processing steps used well-established and trusted methods. These included:

- cleaning: there were some errors in the way people responded, including invalid or inconsistent answers to questions and incomplete responses; online collection methods minimised these errors, but we still conducted checks to detect and resolve them
- resolving multiple responses: the collection design made it easier to submit multiple responses for the same person – for instance, if a family made one response, but an individual also chose to make a separate individual response, cases were detected and resolved within the processing stage
- imputing items: having identified missing, invalid or incomplete answers within submitted responses, we used standard, externally assured statistical approaches to impute the responses robustly, to give a complete and consistent dataset
- estimating and adjusting for non-response and duplicate responses using the Census Coverage Survey (CCS) and other methods: our published census statistics are estimates of the entire population, making an allowance for people who did not respond
- validating resulting estimates against comparator data sources to evaluate their accuracy

Further information on these processes is set out in our Design for Census 2021 article. A summary of these processes is provided in the following sections.

4 . Cleaning the data

Once collected, census data records were passed through a validation and cleaning process. This involved:

- removing duplicate responses
- removing invalid records and responses
- imputing responses to mandatory questions where they have not been completed

These are standard data-cleaning processes used in the production of most official statistics. The following section describes findings from these processes.

Duplicates and false persons

The digital-first design of Census 2021, where most returns were completed online, led to a higher number of duplicate individual and household returns than in 2011. Separate electronic and paper questionnaires were sometimes submitted by different household members. Our Resolve Multiple Responses (RMR) process detected and resolved these duplicate submissions into a single coherent response.

Some submitted responses can contain so little information that it becomes difficult to determine whether the response is genuine. In the absence of some core census variables, these person records (referred to as false persons) are difficult to process and risk creating over-coverage. By making these core questions mandatory in the electronic question, the digital-first census made it harder to submit such false responses, thus improving data quality. The significant reduction in paper responses compared with 2011 also avoided scanning errors affecting overall data quality.

Missing responses to individual questions

Where a response to an individual question is missing, invalid or inconsistent, it is necessary to impute responses. We do this using established standard and assured statistical methods. Responses to voluntary questions are not imputed, as not responding is a valid answer. Far more people completed Census 2021 online, where validation checks were built into the questionnaire. This meant the overall level of item imputation required was lower in 2021 than in 2011.

For example, for the age variable, as derived from the date of birth question, 0.15% were missing a response compared with 0.60% in 2011. For the sex question, 0.27% were missing a response compared with 0.42% in 2011.

We will provide further detail on levels of item missingness in future publications.

Quality control

A comprehensive quality control strategy was designed to check all aspects of the data, with checks conducted only by individuals with suitable security clearance. This strategy was applied at every stage of processing and comprised:

- structural checks ensuring the underlying data retained their overall integrity as they passed through each processing method; this included comparing column and row counts with expected totals and confirming all data values were within expected ranges
- carefully designed diagnostics for all processing methods to monitor performance and check each method was doing what was expected; this provided early insights into the quality of collected data and how people were interacting with the census questionnaires
- ongoing quality assurance and data quality monitoring against a detailed set of planned diagnostics; these were based on thresholds observed in 2011 Census data and topics most likely to be of interest to users, while also including bespoke investigations into emerging potential data issues

Information on these quality control steps was shared and discussed through daily quality management meetings. These were attended by relevant experts and representatives from teams involved in census data processing and analysis. They were also supported by a formal governance and decision-making process.

The quality control strategy helped us to identify potential data quality issues as early as possible. This also helped to define and use appropriate interventions quickly and efficiently where necessary.

5 . Using estimation to assess response

Census estimation is the process of producing population estimates from the data collected. It builds on the data cleaning and imputation in earlier steps, and uses a statistical technique called Dual-System Estimation (DSE) to assess response. More information on the DSE method can be found in Trout, Catfish and Roach: guide to census population estimates (PDF, 817.1KB).

The main aspects of estimation include:

- under-coverage estimation for households
- under-coverage estimation for persons in households
- under-coverage estimation for persons in small communal establishments (see later for large communal establishments)
- over-coverage estimation for those who are duplicated or counted in the wrong place

Each of these stages is complex and made up of several processes. They are based on building national logistic regression models using the Census Coverage Survey (CCS) matched with Census 2021 data.

Census Coverage Survey

The CCS is an important tool for producing census population estimates. The CCS consisted of short interviews with every household in a random sample of unit postcodes. CCS interviewers identified all households in the sampled postcodes and interviewed them. The total sample was approximately 350,000 households.

We matched the CCS to collected census data using standard matching methods and used rigorous clerical checking to maximise the quality of data linkage. Using the matched data, we modelled how likely a given person or household was to respond to the census (or respond incorrectly in the case of over-coverage). We used this information to produce non-response and over-coverage weights, which were used to adjust counts in the collected data to produce a final estimated population total.

For over-coverage, we estimate how likely someone is to have filled in the census twice, or to have been recorded at the wrong address. For example, a student being counted at their out-of-term address rather than their usual address. We also use census-to-census matching to identify the duplication rate more exactly, as census is larger than the CCS.

Our target CCS interview rate was 90%, while the achieved rate was 61%. Although the estimation methods work best when response to both the census and CCS are high, they still work well when only one falls below its target response level. This is especially true when the census response is very high, which in this case it was.

Our judgement, supported by external assurance, was that the methodology could still produce high-quality statistics with the achieved CCS response. The improvement from moving to a national statistical modelling approach gave additional protection. However, it was recognised there may be a widening of confidence intervals. Pre-planned quality assurance and bias adjustment methods became more important, providing assurance that any issues were detected and corrected in the final estimates. We explore this in more detail later.

Response levels in large communal establishments

The coverage estimation modelling approach explained previously is suitable for people in households and smaller communal establishments. However, it is not feasible to run an interview-based survey like the CCS in large (50 or more residents) establishments.

Our large communal establishment estimation strategy involved comparisons between census responses and administrative data sources. By reconciling differences between these sources, we calculated census under-coverage by age and sex for each individual large establishment. This approach was an improvement from the 2011 approach, as data on age and sex were not available from many 2011 administrative data sources.

We received robust administrative data from:

- the National Health Service (NHS)
- Home Office
- Ministry of Defence
- Ministry of Justice
- United States armed forces

We also conducted special surveys for main establishment types, to get their own administrative data on their usually resident population. We had a strong response from these surveys:

- 92% for university owned halls of residence
- 81% for privately run halls of residence
- 92% for boarding schools

Further information on response levels in large communal establishments will follow in a future publication.

Estimation bias and the Alternative Household Estimate

Bias in estimates can be caused by something systematic leading to estimates being unrepresentative, for example, non-response from a certain population group. Our aim is to produce unbiased estimates. This is the principle behind the standard design and why we have our optimising response strategy: to ensure we get adequate representation in the census and CCS.

We created the Alternative Household Estimate (AHE) to help assess any biases in the population estimates. Bias could occur when, for instance, a household refuses to respond to the census and CCS, or in how a household that responded to the census is also more likely to respond to the CCS.

We assessed for potential bias that could result from assumptions underpinning our statistical methods not holding. We did this by comparing household DSEs against the AHE, as an independent estimate of occupied households. The AHE was created using information from responses, field observations and other census collection information, and a range of administrative sources available for households.

Further information on the AHE will follow in future publications.

Difference between return and response rates

Return rates are our operational measure during the collection phase. They are calculated by dividing the count of unprocessed household census form responses received by the count we expected to receive, as per our collection address frame. We use these to ensure we give sufficient attention to each area to achieve our aim of consistent coverage and quality across areas. Return rates are only calculated for households.

Response rates are our output measure. They are calculated by dividing the count of sufficiently complete responses by the number we estimated should have responded. We use these to produce a measure of observed characteristics we collected, which we can compare with our estimate of the true population. Where our Census 2021 response rate was 97%, this means we estimated 3% of the population. Response rates can be calculated for both persons and households.

Figure 1 shows the number of local authorities by person response rate for 2011 and 2021, demonstrating the improvement in response rates when compared with 2011.

Figure 1: Person response rates: count of local authorities in England and Wales, 2011 and 2021

Person response rates for Census 2021 were higher across local authorities compared with the 2011 Census

Download the data

.xlsx

We produce return and response rates for several reasons. They:

- help users to understand the coverage and quality of census data
- provide confidence we have maximised response to the census
- enable us to see what proportion of each census measure is the result of observed data
- assure us there are no gaps in the population estimates for specific groups

Response rates for individual LAs can be found in our local authority comparison tool.

We will publish further information on these measures, and how they are used to calculate confidence intervals, later this year.

6 . Data validation

Data validation is the evaluation of the accuracy of census estimates. Several tools and methods are used to validate the estimates against a range of comparator sources. This identified areas, population groups and topics where there were inconsistencies requiring further investigation.

How we validated Census 2021 data

Our validation of census population estimates involved demographic analyses (for example, looking at the sex and age structure of populations within each local authority (LA)) and comparisons with other sources. These included:

- mid-year estimates (MYE) of population rolled forward from the 2011 Census
- administrative data-based population estimates (ABPE) produced by the Office for National Statistics (ONS) independently from the 2011 Census using linked administrative records
- administrative and survey sources

Where inconsistencies were observed for LAs, we explored these further at small area level. Further information on the sources used in validation are summarised in our Approach and processes for assuring the quality of the 2021 Census data article.

In developing the data validation strategy, we asked LAs which sources they would use to give them confidence in the estimates. Council Tax data (with exemptions or discounts) were identified as an important source to indicate the number of addresses likely to be occupied. Most local authorities provided us with anonymised and strictly controlled Council Tax data for their areas, which were widely used in quality assurance processes.

Our digital-first census design allowed us to identify, investigate and act on anomalies early. Using internal expertise and topic knowledge on subjects such as demography, housing, labour market and health, we assessed emerging findings shown by the provisional census results in the context of other evidence and trends.

External assurance was also sought and provided by, for example, the Migration Statistics Expert Group, the Methodological Assurance Review Panel and a final Executive Assurance Review panel, consisting of the National Statistician and Chief Statistician for Wales.

For the first time we also invited local authorities to assist in the quality assurance of provisional data for their local area.

Local authority involvement in data validation and quality assurance

We worked with LAs to develop our validation plans before and during the collection of census data itself. LAs offer unique expertise and local insight, and we were pleased to collaborate with local government during the validation and quality assurance process.

We offered LAs, county councils and combined authorities the opportunity for controlled early access to indicative rounded census estimates, in line with the Code of Practice for Statistics.

Representatives were asked to compare and identify inconsistencies between the shared rounded census estimates and any locally held sources or intelligence. This feedback was used in validation processes, alongside the investigations already under way.

The rounded census estimates shared included:

- population by single year of age and sex at LA level
- population total at Lower layer Super Output Area (LSOA) level
- full-time students at LA level
- short-term residents at LA level
- population by selected country of birth at LA level
- occupied households at LSOA level
- usual residents by broad type of communal establishment at LA level

In total, 255 organisations (with 554 representatives) took part. Feedback was received from 172 of these organisations (including feedback stating no inconsistencies were being raised).

A range of local insight was provided as feedback. This included many sources we were already investigating in parallel as part of our validation. This feedback proved invaluable in helping to:

- inform the quality of census data
- put the data into local context
- assist in identifying where adjustments to our estimation approach were required

We will publish further information on the investigations carried out in a wider overview of our quality assurance and validation work in a future publication.

Census Quality Survey

We conducted the small-scale Census Quality Survey (CQS) shortly after Census 2021. This was to help us further evaluate census data accuracy. By comparing responses from the census with those provided to the CQS, we can understand better how people completed the census questionnaire. Future analyses will investigate the level of consistency between the two.

Further information on data validation, including the sources available and results of CQS analyses, will be published later in the year.

7 . Changes made because of the quality assurance process

The high return rate, coupled with high-quality responses, meant our approach worked as intended overall. This was thanks to the processes we put in place.

During our quality assurance stages we identified several issues, which we addressed by adjusting the approach set out in our initial statistical design. Local authority (LA) feedback was essential to focus and corroborate across the range of information we considered as part of our quality assurance process.

Adaptations to the design were made in student enumeration and processing, and coverage estimation processes.

Changes to student enumeration and processing

A student's "usual residence" is their term-time address rather than where they happen to be on Census Day, as defined by our residential address and population definitions for Census 2021 article. Government guidance issued as part of the third national lockdown meant some students had not returned to their term-time accommodation after the Christmas holidays by March 2021.

Prior to the census, we published our plans on how we would ensure an accurate estimate of students. In line with these plans, several steps were taken to ensure we could produce statistics on students who still had a term-time address in March 2021. These steps were:

- to provide students with instructions on how they should complete the census if they had a term-time address, through their university and higher education providers
- where a student stated they had a term-time address on the census form of their parents' (or another out of term-time) address, we introduced a step to copy their data to the provided term-time address if no response had been received from them at that address
- to carry out a student hall survey with university accommodation providers shortly after the census collection operation to determine the number, and age and sex profile of students who had a contract to live in a hall in March 2021
- to carry out data linkage work to better understand how to quality assure census estimates against administrative data; this was published in our Understanding students across administrative data in England and Wales article

Changes to coverage estimation processes

Higher level of uncertainty (variance) in estimates

This was identified in feedback provided by local authorities (LAs) who noted differences with occupied Council Tax data, which could not be explained by more detailed investigations.

As mentioned previously, we used the Census Coverage Survey (CCS) to estimate the total population. The CCS performed better in some areas than others. In LAs where the CCS did least well, the estimates had a higher level of uncertainty indicated by estimates of variance from provisional confidence intervals.

We then compared with our independent Alternative Household Estimate (AHE). LAs needing adjustment were identified as outliers based on the distribution of variance across all LAs. An LA was an outlier if their variance was larger than 1.5 times the inter-quartile range, a standard outlier identification approach.

There were four LAs where the variance was high and where the estimated number of occupied households was lower than the AHE. These were:

- Burnley
- Harlow
- Newport (refer to "Newport and Powys" in this section)
- Westminster

In these cases, we calibrated the estimated number of households up to the AHE (with person estimates also changed as a result). The potential use of the AHE as a minimum constraint in this way had been developed as part of our standard design.

There were 15 LAs with high variance and where the estimated number of occupied households was higher than the AHE. These were:

- Barking and Dagenham
- Conwy
- Gateshead
- Gwynedd
- North East Lincolnshire
- Powys (refer to "Newport and Powys" in this section)
- Rother
- Rotherham
- Runnymede
- Somerset West and Taunton
- Stroud
- Thanet
- Tower Hamlets
- Waltham Forest
- West Northamptonshire

Calibrating down to the AHE tended to result in implausible numbers of estimated households compared with estimated persons.

Our approach was to remove the random effects element of the coverage model for these areas, and to allow the national model to predict response. The random effects element is specific to that local area and reflect localised CCS difficulties. Removing this part of the model made the estimates more plausible when compared with our validation sources in our quality assurance and less variable.

Newport and Powys

In December 2022, we announced a processing error for Newport and Powys in our using the alternative household estimate (AHE) with Census 2021 methodology. This meant that the adjustments described above were not correctly applied for these two local authorities.

Estimates for Newport should have been produced by constraining the number of households to the AHE. This was not done. Correcting this error would mean that that the estimated population of Newport would be 128 (0.08%) higher than the published census figure.

Estimates for Powys should have been produced by removing the random effects element of the coverage. While this was done, a further step of constraining the number of households to the AHE was also, incorrectly, applied. Correcting this error would mean that the estimated population of Powys would be 276 (0.21%) higher than the published census figure.

The impact of the error is small in the context of other sources of uncertainty around the estimates and we judged that the benefits to users of continuing with the planned publication schedule outweighed the benefits of delaying those publications to correct the figures.

Bias in estimation

This was identified in LA feedback with differences between census estimates for children and School Census data, as well as differences in the number of occupied households.

During our validation stage, we used LA feedback and compared with our AHE. We found one area (Middlesbrough) was implausibly low in the context of differences seen in other LAs. It was unusual as it was an outlier against both comparators, which we took as evidence of potential bias.

In this area we increased the estimated number of occupied households to be equal to the AHE, similar to our action on high-variance LAs.

Person response rate higher than household response rate

The coverage assessment process used a modelling approach to estimate the level of non-response. It did this separately for people in households, and for the number of households themselves. For 10 areas, the outcome of the models implied there were more occupied households missing than people missing. These were:

- Babergh
- Braintree
- Charnwood
- Chorley
- Fareham
- Hambleton
- Mid Sussex
- Powys
- Rossendale
- South Lakeland

We would not impute a household containing no people, so for those areas we calibrated the person response rate to the household response rate.

Low estimated number of babies and very young children

There was a low estimated number of babies and very young children. This was identified in LA feedback, which noted differences in birth registration data and the number of babies and very young children on NHS data.

Our validation was corroborated with LA feedback across the whole of England and Wales for ages 0 to 2 years. We used estimates for the total population at each age derived from births and deaths data and calibrated each LA estimate up to that total. To avoid overestimation in individual areas, NHS Personal Demographic Service data were used as an upper limit.

Low estimated number of 3- to 15-year-olds

There was a low estimated number of 3- to 15-year-olds identified by the Welsh Government when referencing against School Census data and other sources for children who would not be included on the School Census.

Estimates for Wales were calibrated to a combination of sources on state, independent and home-schooled children, with a similar adjustment for 3- and 4-year-olds to avoid a discontinuity. Analysis of census estimates against School Census data by English region showed the North East was a similar outlier. Adjustment factors calculated for Wales were also applied to the North East, with changes applied to all LAs in these areas.

8 . Future developments

We will publish further articles relating to census quality and methodology later in 2022. This will include publications regarding our:

- full quality assurance assessment of Census 2021 population estimates, including the local authority feedback process
- coverage assessment and adjustment processes
- Alternative Household Estimate and large communal establishment adjustment processes
- data capture, coding and cleaning processes
- non-response, edit and imputation strategy

9 . Related links

Compare age-sex estimates from Census 2021 to areas within England and Wales
Article | Released 28 June 2022
An interactive tool to compare local authorities in England and Wales using age-sex estimates and quality assurance information.

Quality and methodology information (QMI) for Census 2021
Methodology | Released 28 June 2022
Details the strengths, limitations, uses, users and methods used for Census 2021, England and Wales.