Using the alternative household estimate (AHE) with Census 2021

Methodology for the alternative household estimate and how we used it to validate final dual system estimation (DSE) estimates.

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

Release date:
9 December 2022

Next release:
To be announced

Table of contents

1 . Main points

- The alternative household estimate (AHE) provides an independent alternative approach to estimating the number of occupied households for Census 2021.
- It utilises census responses alongside other data sources to calculate a probability of occupancy for each non-responding address.
- The AHE was used to increase the estimated number of occupied households from the dual system estimation (DSE) in just over 1% of local authorities (LAs).
- This increased the estimated number of occupied households by 3,200 in total, with the percentage increase attributed to the AHE of 0.01%.
- The AHE was also used as an important validation tool to compare against the household estimates; it helped inform decisions and aided in quality assuring the estimates as well as being a useful comparator during discussions with LAs.

2 . The Census 2021 coverage strategy

The census coverage strategy is based on statistical methods that make assumptions. For 2021, the estimation approach used the Census Coverage Survey (CCS), the census responses, and recognised statistical methods to determine the number of people and households missed or captured incorrectly. The Office for National Statistics (ONS) checks the assumptions and the resulting estimates by comparing with other data sources and makes corrections where necessary. Further detail on the differences between the statistical design in 2011 compared with 2021 can be seen in our Design for Census 2021 report.

Under-coverage (when someone is missed) and over-coverage (where someone is counted twice or in the wrong place) are estimated through dual system estimation (DSE). However, an important assumption of the DSE is that the probability of a household responding to the CCS is independent of the probability that the household responds to the census. If this assumption is not upheld, then this can cause dependence bias in the population estimates. Therefore, as part of the 2021 Census Statistical Design (PDF, 583KB) presented to the Methodological Assurance Research Panel (MARP), we created an alternative household estimate (AHE) (PDF, 578KB) as part of the overall coverage strategy.

3 . Using the alternative household estimate

We used the alternative household estimate (AHE) to help validate the final dual system estimation (DSE) estimates, and it was a valuable comparator when making decisions around adjustment. The original intended use of the AHE was as a lower bound for the estimates if there was an indication of dependence bias. Where the estimated number of occupied households was lower than the AHE, then the estimates would be adjusted up to align with the AHE figures. Ultimately, there was only a need to apply the AHE because of dependence bias in one area, which was Middlesborough. The AHE was compared with alternative sources as well as the DSE to validate the estimated number of occupied households. In-depth analysis of the census estimates, comparisons with other data sources, and local authority feedback demonstrated that the AHE could also be used when uncertainty in the estimates was high.

As outlined in our Maximising the quality of Census 2021 population estimates report, for those outliers where the estimated number of households was lower than the AHE, the household-level DSEs were calibrated to the AHE, which translated into an adjustment to the person estimates. This use of the AHE as a minimum constraint had been developed as part of our standard design. Further information on how the AHE was applied is outlined in "How the AHE was applied" in Section 5.

Calibrating down to an AHE value had not been part of the standard design. For the local authorities (LAs) with high variance, but with the estimated number of households higher than the AHE, it was found that calibrating down to the AHE tended to result in implausible numbers of estimated households compared with estimated persons. So, although the AHE helped to validate that the estimates for these areas were not plausible, it was not used directly to change their estimates. Instead, the random effects element was removed from the coverage model, allowing the national model to predict response. The coverage estimation process is described in more detail in our Coverage Estimation for Census 2021 in England and Wales methodology.

4 . What the calculated alternative household estimate looked like

The alternative household estimate (AHE) is made up of predominantly responding addresses. Only 6% of the total number of addresses in England and Wales were non-responding, and the probability of occupancy was calculated and applied through the method further explained in this section. The AHE made use of alternative sources of data to provide a probability of occupancy at each address without a valid census response.

Table 1: Response breakdown

Element	Total	Percentage	Occupied	Unoccupied
Census address frame	26,388,300			
Responses: occupied		91%	24,024,600	
Responses: unoccupied		3%		698,200
Non-responses	1,665,600	6%		

Source: Office for National Statistics

Non-responses were split into the following groups, with different estimation methods.

Final field information: trusted source

Towards the end of the census collection operation, field officers made final visits to non-responding households to make a final judgement on whether each was occupied or vacant. Where the field officer got this information from a "trusted source" (the householder themselves, or a trusted third party like a neighbour or gatekeeper), we were able to use that information directly to determine whether the household was occupied or vacant. In addition, if field officers made more than four visits to an address and there was no sign of occupation at each visit, we also considered this a trusted source.

Calculated using final field information and administrative data indicators

However, as set out in section 3.2.2 of the 2011 Census Household Bias Adjustment (PDF, 98.5KB) publication, for the remainder, we could see that where we had final field observations but also had a response, there was a field misclassification rate. There were cases where the field observation said a household was occupied, but we had received a response confirming no-one usually lived there, and vice versa.

For these cases, we calculated the probability that the household was occupied by making use of those cases with a response and a final field observation, and other indicators of occupancy or vacancy in administrative data sources. Further detail on the data sources and method used is available later in Section 5.

Visitors' usual residence

We separated out non-responding addresses for which we had additional information from another census response, where people visiting another address had identified this was their usual address. We did a similar calculation of occupation probability, separating out these addresses because they were shown to have a much higher probability of occupancy.

Identified as holiday homes

Similarly, where someone had identified on a census return that they had an alternative address that was their holiday home, and we had not received a separate response for that address, we calculated the occupation probability. These were shown to have a much lower probability of occupancy.

Calculated using partial field information and administrative data indicators

A subset of non-responding addresses did not get the final field observation, for one of two main reasons. Firstly, this happened if the field operation had already de-prioritised the address as being most likely vacant. Secondly, this occurred when field officers were moved to higher priority areas that had not reached their targets, if the immediate area had already hit return-rate targets. This happened because of the need to get a balanced return rate across all local areas.

In these cases, we were able to use information from previous field visits and administrative data indicators, and we used a weighted-average approach using the field mis-classification rates from the other households that did get the final field visits.

Table 1 shows the starting numbers of responses and non-responses. Table 2 then breaks down the non-responding numbers into the estimated number of occupied households as a result of the AHE.

Table 2: Resultant alternative household estimate breakdown

Element	Total	Percentage	Occupied	Unoccupied
Census address frame	26,388,300			
Responses: occupied		91%	24,024,600	
Responses: unoccupied		3%		698,200
Non-response: final field information, trusted source	729,000	44%	96,500	632,532
Non-response: calculated using final field information and administrative indicators	761,300	46%	555,000	206,200
Non-response: visitors' usual residence	10,300	1%	9,000	1,300
Non-response: identified as holiday homes	4,600	<1%	200	4,400
Non-response: calculated using partial field information and administrative indicators	160,400	10%	69,100	91,300
Total address frame, whether occupied or vacant			24,754,400 -94%	1,633,900 -6%

Source: Office for National Statistics

Table 3 demonstrates that, of the occupied households that should have responded, 97% of them did respond. This aligns well with the final household estimates produced through our coverage-estimation approach.

Table 3: Resultant household response rate according to the alternative household estimate

Element	Occupied	Percentage
Census occupied addresses	24,754,400	100%
Responded	24,024,600	97%
Non-responses	729,800	3%

Source: Office for National Statistics

5 . Methodology: calculation and application of the alternative household estimate

The alternative household estimate (AHE) methodology can be broken down into four clear processes: the creation of the address-level dataset, processing the underlying data, calculating the probability of occupancy, and applying the AHE in the production of the final estimates. Each process is outlined in this section under its own subheading.

The methodology for creating the AHE was discussed at the Methodological Assurance Research Panel (MARP) in October 2021, as described in the Alternative Household Estimate 2021 (PDF, 578KB) report. Following the discussion with MARP, we applied the original methodology to the live data and made additional improvements to the AHE during multiple thorough reviews of the AHE estimates. We discussed these suggested improvements with MARP.

Creating the address-level dataset

To create the AHE, multiple data sources were brought together using linkage of their Unique Property Reference Number (UPRN). These sources were:

- census response data
- census management information (MI)
- Personal Demographic Service (PDS) data
- English School Census (ESC)
- Council Tax
- utilities data
- Valuation Office Agency (VOA) data
- data from housing association authorities confirming an address is vacant

Census MI includes data on interactions by an address with the census process, for example a request for a paper form to be sent, and data from field officers. The data from field officers include the outcome of each visit made by field officers during the non-response follow-up period and information on the final field outcome, for example, the address is occupied because of a neighbour saying someone lives there (termed a "dummy form").

This was a great improvement from the method in 2011, where the AHE was based on dummy form information from only the Census Coverage Survey (CCS) areas (which used a 1% sample of postcodes). Having additional sources of data, which were all at address level, meant more granular data and stronger indications of occupancy as a result of multiple indicators. One such improvement was that all MI from field officers was electronic, which meant data were easy to access and use for these purposes. In 2011, all field MI was handwritten and had to be manually keyed in for use.

From the combined data, some addresses are removed, for example, communal establishments or addresses marked as invalid by field. This was to ensure that the AHE only included addresses that were households that should have responded to the census.

Processing the underlying data

Once we had filtered the combined addresses, we then added a set of derived variables. These were:

- to indicate occupancy from census response data
- to indicate occupancy from field officer dummy form data
- two indicators of occupancy or vacancy from alternative sources

How these are derived is outlined in the following categories.

The census response indicator is derived as follows:

- where a census response confirms someone lives there, this is considered "occupied"
- where a census response confirms no one usually lives there, this is considered "vacant"
- where no response has been received, this is considered "non-responding"

The field officer dummy variable is derived as follows:

- an address is considered "occupied" if the field officer noted it as an absent household, extraordinary refusal, or hard refusal
- an address is considered "vacant" if the field officer noted it as holiday accommodation, a second residence, or vacant
- where the field officer could not make contact, the address is noted as "no contact"
- where the field officer did not register a final visit outcome, the address is noted as an "unaccounted for address" (UFA)

Where any of the following indicators were present, the occupied indicator for that address was classed as "occupied". If there were no instances of these indicators across the sources, then the address was classed as "unknown".

Alternative sources for occupied indicators

- Census 2021 responses: an address had been identified as a usual residence of a visitor in another census response and was therefore occupied.
- English School Census: if the count of children at an address was greater than zero on the English School Census, then it was believed the address was occupied by usual residents.
- Census MI: if there was at least one request for a household paper form to be sent to the address, it indicated that someone would have been at the address to receive the paper form.
- Census MI: any indication from the field officer that indicated the address was occupied, for example, a request to call back another time or a hard refusal.
- Personal Demographic Service (PDS): where a PDS record had been recently updated (within three months up to Census Day) and there were people registered at that address.
- Utilities: an address had been identified as having a good indication that it was occupied through gas or electricity consumption.

Where any of the following indicators were present, the occupied indicator for that address was classed as "vacant". If there were no instances of these indicators across the sources, then the address was classed as "unknown".

Alternative sources for vacant indicators

- Council Tax: a discount or exemption indicates the address was an empty, unoccupied, or second home.
- Census 2021 response: another census response indicates the address had been identified as a holiday home and was therefore empty.
- Housing association data: the address has been confirmed vacant through the landlords of housing association addresses.
- Utilities: an address had been identified as having a good indication that it was vacant through lack of electricity consumption.
- Census MI: any indication from the field officer that indicated the address was vacant, for example, a visit concluded the address was not occupied.

Determining the probability of occupancy at each non-responding address

The majority of the addresses were then split between responding addresses and non-responding addresses.

The responding addresses remained unchanged and therefore made up a large part of the AHE figures. If a census response stated that the address was occupied or vacant, it remained as such. The probability of occupancy calculated was only applied to non-responding addresses. The count of responding addresses and non-responding addresses is shown in Table 1.

Deterministic rules: final field observations and trusted sources

While developing the AHE, some of the non-responding addresses had clear indications that they should have been set to "definitely occupied" (a probability of occupancy equal to one) or "vacant" (a probability of occupancy equal to zero). These addresses were separated from the rest and had specific deterministic rules applied to them. It is worth noting that, within the deterministic rules, there are responding addresses that appear, but the census responses for these addresses are not altered. These addresses continue to be occupied or vacant as the response confirmed. The following information outlines these addresses and how they were treated.

Deterministic rules were set for specific groups of addresses. The following types of addresses were set to vacant:

- caravan park addresses originally marked as vacant by field officers because they were observed to be locked shut during the coronavirus (COVID-19) pandemic
- housing association properties that the local housing associations advised were vacant as at Census Day
- addresses where there had been more than four field visits, and each of these gave no indication that the address was occupied, and the final dummy completion was "vacant"
- caravans and other mobile structures, where there was no response and the dummy form indicator was "vacant", because there was a much higher probability that these would not have contained usual residents
- addresses with no final field information, but previous visits and other sources implied they were either vacant or did not give any confirmatory signs of occupancy
- addresses where the dummy form indicator was "vacant", and the source of that information was from a "trusted third party", for example a gatekeeper, neighbour, or the householder themselves

Further to the last rule, there was also a deterministic rule where addresses were set to occupied if the dummy form indicator was "occupied" and the source of that information was from a "trusted third party".

We applied these specific deterministic rules after multiple iterations of checks against the DSE and other comparative data sources. There was a strong justification for each rule, and the findings when applied improved the final estimate from the AHE. The field officer visit information was highly valuable intelligence on what they had seen during census collection, which proved incredibly useful in making decisions on some addresses and most of the deterministic rules.

Calculating the probability of occupancy for all remaining addresses

The majority of the remaining non-responding addresses had a probability of occupancy calculated based on responding addresses.

All addresses that did not match any of the deterministic rules described were split into strata, by:

- similar LA group
- field officer dummy form outcome
- alternative source indicator of occupancy
- alternative source indicator of vacancy

It may be expected that the alternative data sources should be used to suggest either an occupied or vacant address, however, previous experience showed that we could not trust just one source of information on its own. In 2011, we examined the dummy form information from responding addresses, as set out in section 3.2.2 in the 2011 Census Household Bias Adjustment (PDF, 98.5KB) publication. We found that dummy form indicators did not always align with what the census response had shown. This check was replicated in 2021 for all additional sources and found similar results. Therefore, by creating strata that combine multiple different sources of information and calculating the agreement rate with those addresses that have a response (which says the address is either occupied or vacant), all data sources are considered when calculating the final estimate. These calculations gave an occupancy rate, which could be applied to any non-responding address in the same stratum, for example, with the same dummy and administrative data indicators.

Rather than calculating at LA level, we grouped the data by the 2011 Census similar LA groupings (XLS, 486KB) to ensure there were sufficient observations within each stratum to provide robust calculations to be applied to the non-responding addresses. We calculated the occupancy rate for visitors' usual residences and respondents' holiday homes, as previously described, using all observations for England and Wales, as there were not sufficient cases in each LA group.

An example of how the probability of occupancy would be calculated and applied to a non-responding address in a given stratum is provided in Table 4.

Table 4: An example of the alternative household estimate calculation method

Similar local authority group	Dummy form	Alternative source occupied indicator	Alternative source vacant indicator	Total responding address occupied (A)	Total responding addresses vacant (B)	Total number of responding addresses (C)	Proportion occupied by stratum from responding addresses (A divided by C)
London cosmo-politan	Occupied	Occupied	Vacant	200	100	300	0.67
			Unknown	70	10	80	0.88
		Unknown	Vacant	100	100	200	0.5
			Unknown	50	25	75	0.67

Source: Office for National Statistics

The strata in Table 4 are not exhaustive; there are also strata for each similar LA and all the additional groups not captured in Table 4. For example, there are also the categories "dummy form: vacant" and "dummy form: no contact", and these categories will be broken down by alternative source indicators as well.

For any non-responding addresses in the strata, the final probability of occupancy in the last column will be applied to that address. Therefore, no single indicator provides the probability of occupancy; it is calculated from the number of responding addresses (and whether they are occupied or vacant) across each stratum.

The final probability of occupancy was then grouped into local authority, hard-to-count index, and accommodation type before it could be used in the estimation process. An example of this is shown in Table 5.

Table 5: An example of the final file used to apply the alternative household estimate in the estimation process

Local authority	Hard-to-count index	Accommodation type	Alternative estimate of occupied households (sum of each probability of occupancy for each individual address in the group)
E010000000	1	Detached	300
E010000000	1	Purpose-built flats	200
E010000000	1	Part of converted building	45
E010000000	2	Detached	140
E010000000	2	Purpose-built flats	400
E010000000	2	Part of converted building	10

Source: Office for National Statistics

How the AHE was applied

As previously mentioned, producing population size estimates corrected for coverage error using the Census Coverage Survey and census data assumes independence between these two data sources. Independence means that, for every member of the target population, a chance of responding to the Census Coverage Survey does not depend on the member being a census respondent or non-respondent. In practice, such independence is not fully achievable, and the dependence bias adjustment of the under-coverage adjusted estimates may be needed.

The approach for the dependence bias correction requires alternative high-quality census counts or estimates at some reasonably low-level geography. Similarly to the 2001 and 2011 Censuses, the AHE was used in Census 2021 for England and Wales to adjust for the dependence bias. However, since the coverage estimation in the 2001 and 2011 Censuses used the post-stratified dual system estimator, while the coverage estimation in Census 2021 used the mixed-effects logistic regression approach, the way the adjustment was applied was very different in 2021. In the 2001 and 2011 Censuses, the alternative household estimates were used to compute the alternative odds ratios (the main estimation assumed them to be equal to one), and then we used these odds ratios to correct for the dependence. The logistic-regression-based approach deals primarily with the coverage probabilities, so a more natural way to adjust for the dependence bias in this case is to work out the alternative coverage probabilities at the alternative household estimates stratum, and then use them in adjustment.

There are two main challenges when using the alternative household estimates to correct for the dependence bias. Firstly, the alternative estimates are available at the alternative-household-stratum level, which is local authority by hard-to-count index by accommodation type. While dependence indeed varies by local authority by hard-to-count index by accommodation type, such grouping aggregates households with very different dependence propensities. Therefore, there is a risk that certain types of households will be systematically over-adjusted, while some will be systematically under-adjusted. The second challenge is related to the fact that the reliable alternative estimates are available for the household population only, whereas the dependence bias adjustment is required both for the household and person populations.

There were several dependence bias adjustment methods designed and tested at the research stage for Census 2021 (PDF, 3.2MB). The chosen approach is the direct adjustment method with apportionment. To adjust the household population estimates, the observed census count and the alternative household estimate by local authority, hard-to-count index and accommodation type are used to produce an adjustment weight for the entire stratum.

However, this adjustment weight needs to be applied to households. We allocate this according to the estimated under-coverage by the household structure variable. This variable reflects the broad age-sex grouping, the relationship between the members of a household, and the broad household size. Those household structures that have the lowest estimated response rates are allocated more of the adjustment weight. These apportioned weights are then applied to each original coverage weight within a local authority by hard-to-count index by accommodation type by household structure stratum.

To adjust the person population, the previously described apportioned weights by local authority by hard-to-count index by accommodation type by household structure are re-used to correct individual person coverage probabilities within the stratum.

Impact of the AHE

Once the AHE had been applied in the estimation model, at the national level, the total increase in occupied households attributed to the AHE was 3,200. The number of occupied households added by the AHE by local authority are provided in Table 6.

Table 6: Number of occupied households added to the final estimate by local authority, and the percentage increase attributed to the alternative household estimate (AHE)

Local authority	Number of occupied households added	Percentage increase because of the AHE
Middlesborough	1,431	2.40%
Burnley	510	1.30%
Harlow	271	0.70%
Westminster	1,027	1.10%

Source: Office for National Statistics

Middlesborough's original estimates were adjusted up to the AHE because of an indication of dependence bias, which was the original intended use of the AHE.

During the validation of the original estimates, this was clear when comparing with the AHE and from feedback received from the local authority itself. For all other areas, there was not a strong indicator of bias because of the high census response rate, which highlighted that Middlesborough required some adjustment.

The remaining three local authorities – Burnley, Harlow, and Westminster – were adjusted because of high uncertainty (variance) in the results, causing large confidence intervals. These areas had original estimates that seemed implausibly low and therefore were adjusted up to the AHE. These LAs had been identified as clear outliers, as a result of the variance being greater than 1.5 times the inter-quartile range.

There were other areas with large confidence intervals; however, the original estimate was higher than the AHE, so these were dealt with differently. Our Maximising the quality of Census 2021 population estimates report provides further information on the changes to the coverage estimation process over and above the AHE adjustment.

Overall, the AHE has been heavily improved on since 2011. Equally, the return rates from the census were impressive, and therefore the census estimates did not require as much as adjustment as previously thought. The AHE was therefore only used for its original intention in Middlesborough but proved useful in three other local authorities in providing a good estimate to adjust up to when the confidence intervals from the estimates were large. The AHE has also been a great tool to help in discussions with local authorities, to help provide an additional source of data to compare the estimates with and to validate them. Additionally, it is worth noting that the AHE in all other areas was comfortably within the confidence intervals of the household estimates. This, in turn, demonstrates the robustness of both the estimates and the AHE approach as they reached very similar figures using independent methods.

Newport and Powys

A processing error meant that the adjustments described in our Maximising the quality of Census 2021 population estimates report 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.

6 . Related links

Coverage estimation for Census 2021 in England and Wales
Methodology | Last revised 9 November 2022
Methodology for coverage estimation of Census 2021 in England and Wales.

Maximising the quality of Census 2021 population estimates
Methodology | Last revised 9 December 2022
How we maximised the quality of Census 2021 population estimates during the processing and quality assurance of the final statistics.

7 . Cite this methodology

Office for National Statistics (ONS), released 9 December 2022, ONS website, methodology, Using the alternative household estimate (AHE) with Census 2021