Office for
National Statistics

Coverage estimation for Census 2021 in England and Wales

Methodology for coverage estimation of Census 2021 in England and Wales.

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

Release date:
9 November 2022

Next release:
To be announced

Table of contents

1 . Main points

- Census 2021, as with any census, was subject to non-response and incorrect responses.
- The Office for National Statistics (ONS) developed on our successful approach to estimating coverage error from 2011 and built on the improved processing of Census 2021.
- We used logistic regression models rather than a stratified approach to dual system estimation, allowing us to account for the effect of many more characteristics on response.
- The results then went through a thorough quality assurance process, with adjustments made where necessary; the census count was 97% of the final estimate.

2 . Summary

The Census of England and Wales provides an accurate, comprehensive and consistent picture of the England and Wales population, as laid out in Design for Census 2021. The key aim of the Census is to produce high quality population counts, at subnational and national level, for demographic characteristics, which include:

- age
- sex
- ethnicity
- tenure
- accommodation type
- economic activity

Despite the best effort made by census data collection operations to count everyone, the complexity and size of the population results in census coverage errors. The most prevalent coverage error is when a member of the target population is missed in the census (undercoverage). Less often, but still at a non-ignorable rate, a member of the target population is either duplicated or counted not at their usual residence (overcoverage). The ONS uses a variety of statistical methods to estimate these coverage errors to produce corrected population totals for local authorities by the key demographic characteristics. These estimates in general have higher accuracy than the raw census counts.

The Census Coverage Survey (CCS) was used to estimate the census coverage error. The data from this survey are then linked to the census data. A combination of capture-recapture, analysis of complex survey data, and small area estimation methods are used. Finally, a bias adjustment process is used to adjust for issues which cannot be accounted for in the main design. These can occur as the result of some of the statistical assumptions not being practically attainable in complex data collection exercises like the CCS and census. Variance estimation methods are used to assesses the uncertainty around these estimates.

Census data collection covers both the general household population and managed residential accommodation, known as communal establishments. The general population itself is the population of households and the population of individuals in these households. This methodology focuses on the 2021 Census coverage estimation for the general population, details of coverage estimation for communal establishments will be published in early January 2023.

3 . Overview of census coverage estimation and related processes

The Census Coverage Survey

Similarly, to the 2001 and 2011 Censuses, the Census Coverage Survey (CCS) was used to measure coverage in the 2021 Census outlined in 2011 Census Coverage Survey Summary. Starting eight weeks after Census Day, the coverage survey produces an independent count of the population in a large sample that covers all local authorities in England and Wales. To ensure the operational independence between the coverage survey and census, an independent sampling frame is used for the survey and there are restrictions on interviewers being involved in collecting data for two sources within the same area. The sample contained approximately 16,000 postcodes, which is 1.45% of England and Wales postcodes - these include nearly 340,000 addresses.

The Census coverage survey is a stratified two-stage cluster sample. The sample is stratified by local authority and hard-to-count index. The hard-to-count index has five levels, reflecting the expected census coverage level for a small area Lower Super Output Area (LSOA), approximately 1,500 households. The overall sample size is first allocated to hard-to-count strata using an optimal allocation method. This method puts more sample in strata where lower coverage is expected to mitigate possible increase in uncertainty around estimates in these strata. Therefore, areas that are classified as harder to count are represented disproportionally to their prevalence in the population, as described in 2021 Census coverage survey: sample allocation strategy Burke and Rainskij, 2020. Across the England and Wales population, the split of easiest to hardest to count areas is 40%, 40%, 10%, 8%, 2%. However, within the sample the allocation is approximately 29%, 41%, 12%, 14%, 4% to those same areas.

After this, it is then allocated to local authorities in proportion to their size. The sampling process first selects output areas (OAs), which contain approximately 120 households, within each stratum. Then a quarter of postcodes were sampled from each sampled output area. With a few exceptions, at least two output areas from every local authority by hard-to-count stratum are included in the sample. Within this sample of postcodes, interviewers attempted to enumerate all households within the selected postcodes.

The achieved 2021 Census Coverage Survey's coverage rate was 59%, lower than the target of 90%. The interview completion rate (interviews completed per contacts established) 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.

Removing false persons, editing and imputation, record linkage, and pre-processing

Before the census and coverage survey responses are ready for coverage estimation, there are several crucial processing stages.

The first process, called remove false persons, identified and resolved records where individuals responded to the Census more than once within the same address. Editing and imputation resolved inconsistencies in reported characteristics and dealt with item level missingness by imputing missed values for certain variables. This is outlined in Item editing and imputation process for Census 2021, England and Wales. Note that coverage estimation deals with unit level missingness, when an entire member of the population is missing, while imputation is concerned with responses that have some values missing.

Following the removal of false persons, but alongside the editing and imputation (and therefore using un-edited and un-imputed data) two record linkage exercises take place. First, the coverage survey data is linked to census data (within and outside of the coverage survey postcodes). This linkage exercises establishes whether, within the sampled areas, each survey record has a corresponding census record. That is whether:

- a household or individual responded in both sources
- a survey record has no corresponding census record (such as, a household or individual was missed from census)
- a census record has no corresponding survey record, (such as, a household or individual was missed from the survey)

The results are used to estimate undercoverage. The required quality of linkage for undercoverage is no more than 0.1% of links to be false positives and no more than 0.25% to be false negatives, more information can be found in Methodology report on coverage matching for the 2021 Census. By linking the survey to census outside the sampled areas, census duplicates and returns in the wrong locations can be identified. These results are used to estimate overcoverage. Another linkage exercise, linking the census data to themselves, is conducted to support overcoverage estimation.

Outputs from the editing and imputation processes are then combined with the linkage results. At this point the data are ready for coverage estimation pre-processing.

Estimation received six datasets for Census and five datasets for Census Coverage Survey. These datasets contained all the information needed about the individuals, households, and small communal establishments. Two match lists were also provided with outcomes and information from the matching processes.

Responses that were outside of the target population were marked as out-of-scope. For Estimation, this primarily excludes late returns as we create estimates only on Census cases returned before CCS begins. Then, cases where multiple Census person and/or household were matched to multiple CCS persons or households were then resolved using the following hierarchical logic:

- keep the household match with the best match score given by the Matching team
- keep the household match with the most exact address match
- keep the household match with the most common residents
- keep a match at random

Once a household match was determined, all the other household matches and their resident-matches were marked as out-of-scope.

Once pre-processing was complete, this allowed for the model selection and estimation process to begin.

Census coverage estimation

The census coverage estimation process includes several tasks before the coverage error population total is obtained. It adjusts for undercoverage within the coverage survey sample, then uses this adjusted survey estimate to produce the population level estimate for many small groups like age-sex by local authority and adjust for overcoverage.

Census coverage estimation for the 2001 and 2011 Census of England and Wales

In the 2001 and 2011 Censuses of England and Wales each such task had a separate estimator.

The data from the coverage survey and census within the sampled areas are used to estimate the undercoverage adjusted population size within the areas sampled by CCS. The method is known as dual system estimation. Even though both the CCS and the census are missing some people from the population, under certain assumptions, the ratio of those counted in both sources to those counted in the survey alone is the same as the ratio of those counted in census to the population size.

Dual system estimation works for those areas sampled by CCS, and so the next step was to use the undercoverage adjusted totals at the sample level to produce whole-population estimates. In the census coverage case, the ratio estimator is an efficient estimator that can take the sample estimates and produce the population estimates.

The level at which the undercoverage corrected totals can be reliably obtained with the ratio estimator was not the level required by users, who needed results by age-sex group by local authority. This was because while the coverage survey on its own is very large, it is not always large enough to support the small area estimates. Local fixed effects models were used in 2001 and the synthetic estimator in 2011 outlined in The 2011 Census Coverage Assessment and Adjustment Process.

Undercoverage estimation is followed by the overcoverage estimation. While the extent of overcoverage is clearly non-ignorable, it is quite small and therefore challenging to estimate even with a large survey. Moreover, it is difficult to design a coverage survey so that it simultaneously makes both under- and overcoverage estimation efficient. The design of the coverage survey is focused on undercoverage. In 2011, overcoverage propensity was directly estimated at very high levels of aggregation and combined with the undercoverage adjustment. There was no overcoverage estimation in the 2001 Census.

Census coverage estimation for the 2021 Census of England and Wales

In 2021, a more unified approach based on logistic regression was used for coverage estimation Coverage Estimation Strategy for the 2021 Census of England and Wales. The probabilities of undercoverage and overcoverage errors were modelled using demographic characteristics, geography, field management as well as interactions of some of these variables. These probabilities estimated how likely a member of the population with a certain set of characteristics was to respond to the census or make an incorrect census return.

These probabilities were then transformed and combined to give a weight for each census record based on the characteristics of that record. This weight reflected a net contribution of a person with given combination of characteristics to correct for coverage errors in census. In order to obtain the population size for a group of interest, weights for the group of interest are summed up.

While in the 2001 and 2011 Censuses the success of estimation depended on a careful selection and combination of levels at which different estimators were applied at, the success of the estimation in 2021 depended on careful model selection.

Whenever evidence supported the use of mixed effects logistic regression, this was used. Otherwise, fixed effects logistic regression was used. In the mixed effects model, the local authority was treated as a random effect, while the remaining variables were treated as fixed effects. Mixed effects models allow us to deal with the variability of census coverage at local authority level in a more efficient way than treating it as a fixed effect. Mixed effects logistic modelling was used for the undercoverage estimation of households and persons. However, in overcoverage and communal establishment estimation the fixed effects logistic regression models were fitted and there were no specific effects for local authorities. There were several benefits from moving to the logistic regression approach.

First, because the census data collection in 2001 was quite similar to 2011. This meant the sample in the 2011 Census coverage survey could be allocated very efficiently by using the results of 2001 Census. In other words, we knew where census non-response was likely to be and could allocate the coverage survey there. Moving to the primarily online data collection in 2021 meant increased uncertainty about patterns of census non-response, as it was not clear that online response patterns would be the same. This meant that continuity in the estimation approach was less useful.

On the other hand, online data collection also allowed all the census data to be available for estimation at once, rather than in batches. Fitting a model to the entire coverage survey data does not require such precise knowledge about where to allocate sample compared with the methods used before. However, work on modelling complex survey data in the context of coverage estimation was needed and the sample design was adapted to work seamlessly with the regression approach.

There are also efficiency gains in using the data from entire coverage sample rather estimating for separate population strata or groups. With a carefully selected model the regression approach can achieve smaller overall error for most small areas compared to direct estimation. Most of the research prior to the 2021 Census was done assuming local authority census coverage would vary from around 81% in the hardest to enumerate to 98% in the easiest to enumerate local authorities. Such differential coverage can be reliably reflected in the estimation if a model is selected appropriately.

Carefully controlling multiple variables simultaneously also allows us to better deal with the bias due to pooling individuals with different census response propensities, compared to the post-stratified dual system estimator. That is, we can control for more characteristics, rather than (as in 2011) assuming that response rates are the same for everyone in the same age-sex group in a Local Authority.

Finally, the logistic regression approach can be used both for under- and overcoverage estimation. Assuming a good model selection procedure and processing system, this allows efficient use of resources and faster processing.

Following estimation, a thorough quality assurance process took place, to ensure the estimates were plausible and met user needs. This involved comparing estimates to mid-year population estimates, administrative data sources and feedback from local authorities. As a result, some adjustments were implemented.

In the last step, variance estimation is conducted to quantify the uncertainty around the coverage error adjusted population size estimates.

4 . Census coverage estimation methodology

General population undercoverage estimation and adjustment

Once the coverage survey and census data are ready for estimation and matching complete, within sampled areas we know for each CCS record whether a census response exists or not. This can be used to model the coverage probability given the set of observed census variables (or predictors) and their combinations (interactions).

Logistic regression is a powerful and well-understood tool for modelling probabilities. If modelling is done appropriately, it is possible to increase the precision of estimates thanks to using a large sample. We can also reduce certain errors by simultaneously controlling for several important variables and some of their interactions. The approach uses the entire dataset to relate a combination of demographic and other characteristics to the estimated probability that a member of the population with such characteristics will respond to census. Say, for example, a person who has the characteristics:

- aged 32 years
- male
- white
- living in a rented purpose built flat
- looking for a job
- in a household size of two residents
- related to somebody in the household
- in hard-to-count 3
- in South West of England
- in a self-contained accommodation
- born in the UK
- not a student
- in an area that received access codes
- observed census return rate 0.965

This member of the population would have the estimated census response probability 0.95. While, a person with exactly the same characteristics except being female would have the estimated probability 0.953.

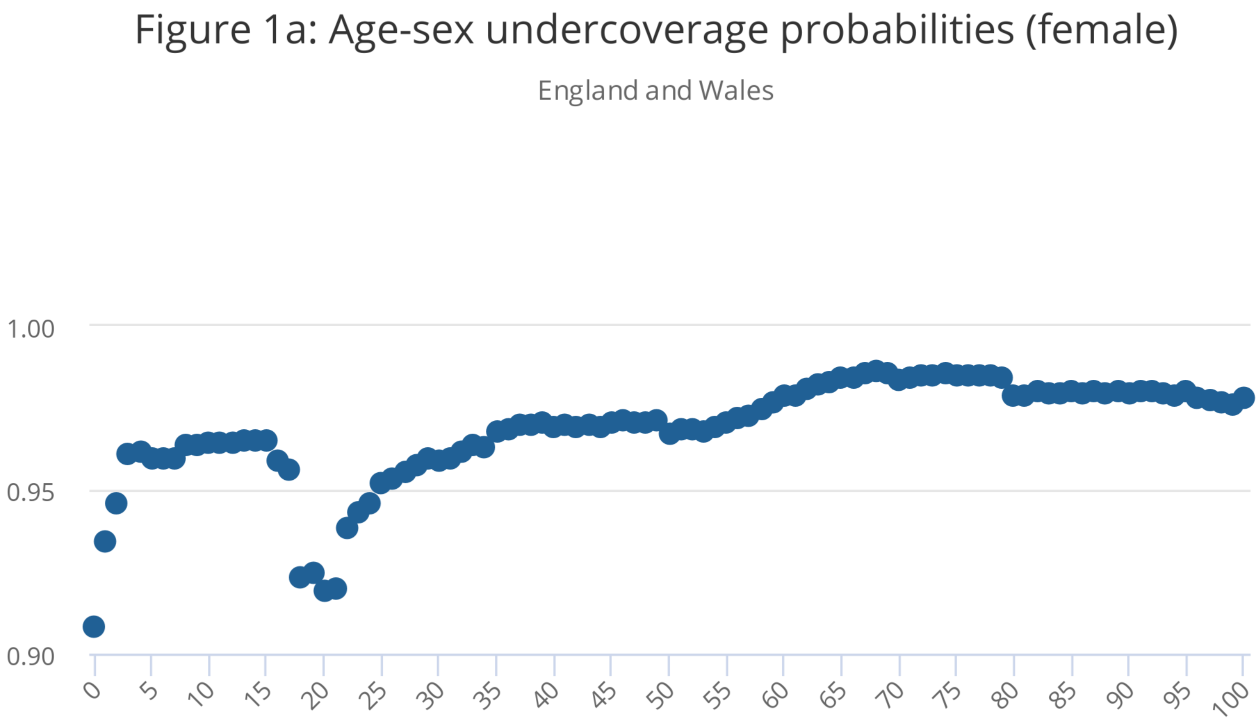
The model predicts the probability for each combination of variables. Therefore, each observed census record gets a corresponding census response probability. This probability can be transformed to a coverage weight by taking a reciprocal of the probability. In the example previously, we will have weights 0.95^{-1} 1.053 and 0.953^{-1} 1.049, respectively. If we observe 1000 individuals in the census data like the first person in the above example, we can sum up the corresponding weights to estimate the population total for individuals with such characteristics: 0.95^{-1}× 1000 1053. Similarly, if we observe 1000 individuals in the census data as the second person, the corresponding estimate is 1049. Of course, we are never interested in such specific set of characteristics, but rather something more useful, say, age-sex group by local authority. Since all census records are weighted, it is possible to produce the undercoverage adjusted total for any group of interest.

Mixed effects logistic regression was used both for household and person undercoverage estimation. It is similar to the model described above, but has the local authority as a "random effect". This random effect allows the model to reflect the area specific variability in a more efficient way than having local authority as a fixed effect (like all the variables in the example shown). Without the random effect the probabilities for these two persons would be 0.95 and 0.953, no matter what local authority within the region these two person were located in. However, with random effects, those probabilities would be, say, 0.962 and 0.964 in local authority A, while 0.934 and 0.938 in local authority B. Mixed effects based estimation reflects local differences, but comes with the cost of increased variability of estimates.

Undercoverage was estimated and corrected for both the household totals and person totals. The general approach was the same in both cases, though the actual models are different. In the case of the household estimation, there was an additional adjustment for the distribution of household size. Model selection for two populations was run independently, but we tried to have as much consistency as possible in terms of levels of variables and interactions used. The two populations are 'reconciled' by the adjustment process, further information on the adjustment process will be published in winter 2022.

Figure 1a: Age-sex undercoverage probabilities (female)

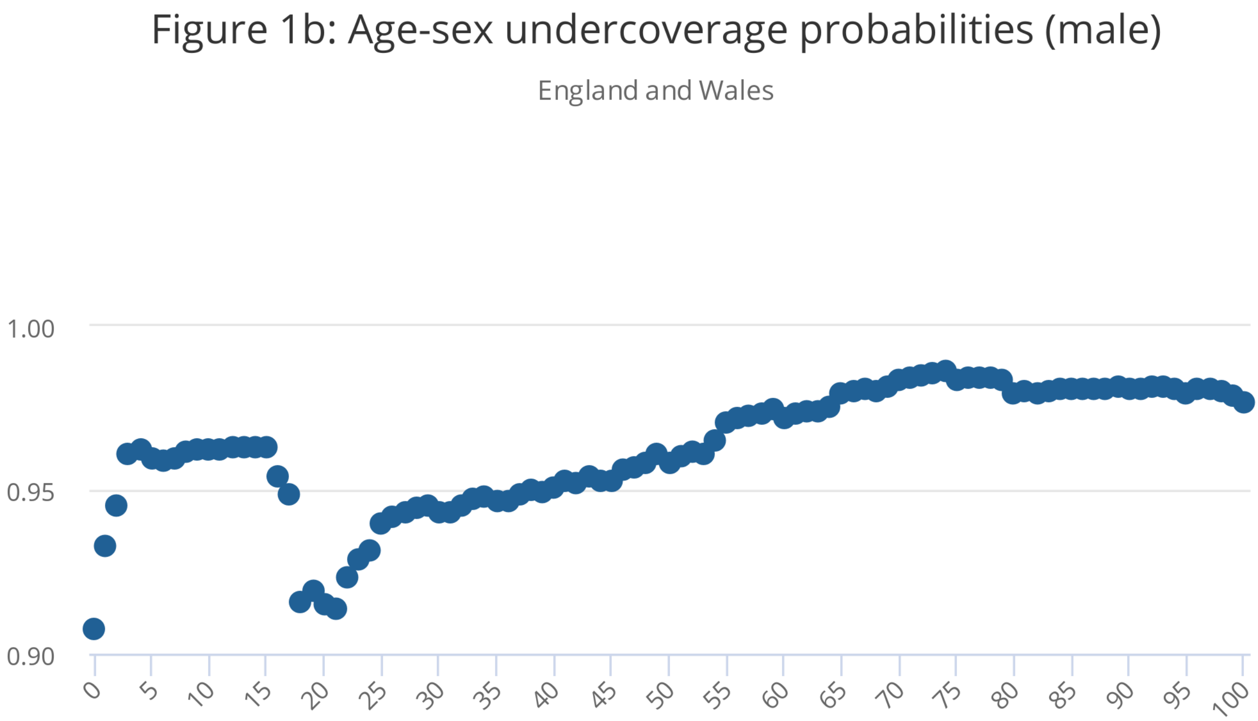
England and Wales



Source: Office for National Statistics, Census 2021

Figure 1b: Age-sex undercoverage probabilities (male)

England and Wales



Source: Office for National Statistics, Census 2021

General population overcoverage estimation and adjustment

A similar approach to undercoverage estimation was used for overcoverage estimation at person level. Overcoverage occurs when a member of the census population is either enumerated:

- more than once
- in the wrong location
- despite not being a member of the target population (e.g. individuals born after census day)
- because of a completely fictitious census return

Where possible, data cleaning resolved erroneous records (Remove False Persons) and multiple responses at the same location (Remove Multiple Responses). As such, in overcoverage estimation we only estimate for individuals enumerated more than once or enumerated in the wrong location.

Instead of modeling the coverage probability of those in the Census, overcoverage estimation was used to estimate the probability of correct enumeration in the census. Much like undercoverage estimation, the linked census and census coverage survey allowed each linked record to have an outcome of 0 or 1, depending on if they were correctly enumerated or not. It is important here to assume there is no overcoverage in the census coverage survey, as it is used as the correct location of census individuals. This is assumed due to the way the census and census coverage survey are designed, where the time between the collection of them is designed to be large enough to optimise response rates but to reduce movement in the population. This linked outcome was used to model the probability of correct enumeration, using a fixed effects, logistic regression model. Both numerical issues in the model fitting process and timescales meant that random effects were not included in this model.

Using the same example of characteristics, that person might have the estimated census correct enumeration probability 0.995. A person with exactly the same characteristics except being female, has the estimated correct enumeration probability 0.9953.

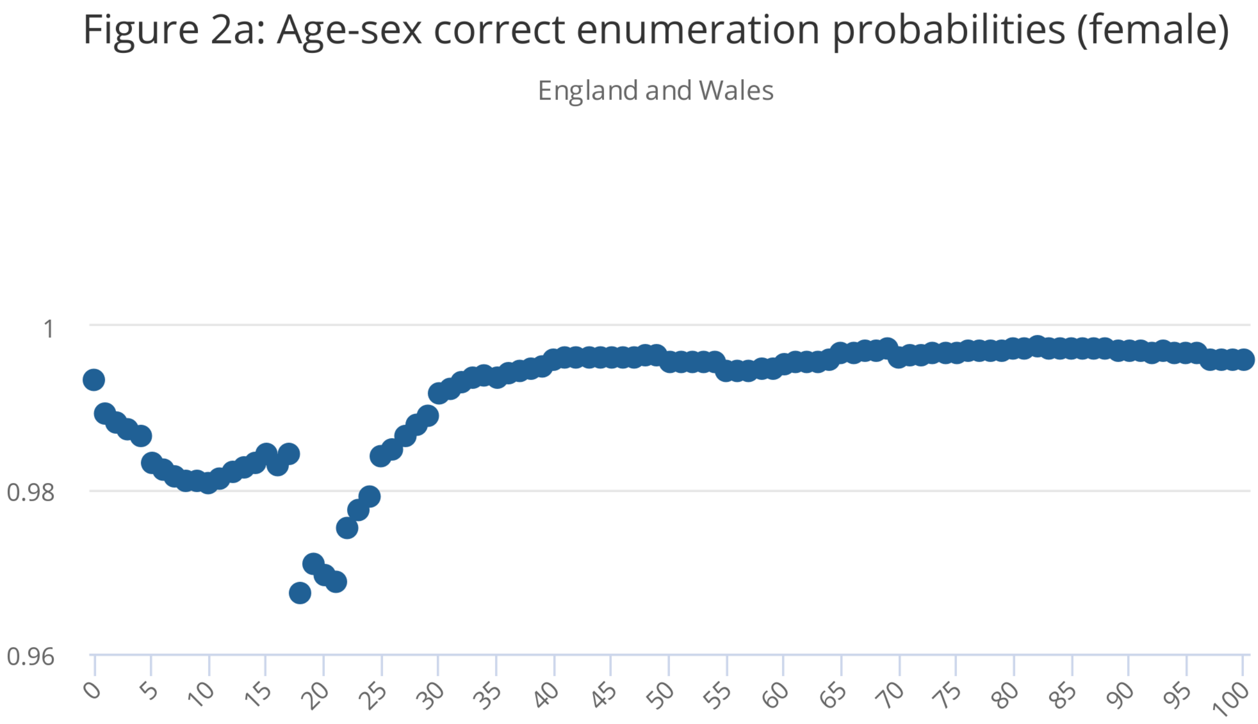
In the same way, this overcoverage model then produces a correct enumeration probability for each combination of variables. Therefore, each observed census record gets the corresponding census response probability and correct enumeration probability. The response probability can be transformed to the coverage weight by taking a reciprocal of the probability. However, for overcoverage estimation, the aim is to down-weight the census estimate and therefore the undercoverage weights are multiplied by the correct enumeration probabilities. Where undercoverage error is estimated and correcting for overcoverge error, we will have weights (0.995x0.95^{-1}) 1.047 and (0.995x0.953^{-1}) 1.044, respectively. If we observe 1000 individuals in the census data, we can sum up the corresponding weights to estimate the population total for individuals with such characteristics: (0.995x0.95^{-1}) x 1000 1047.37. Similarly, if we observe 1000 individuals in the census data as the second person, the corresponding estimate is 1044.

Similarly, to 2011, matching of the census dataset to itself allowed for stronger estimates of the level of duplication, with high precision for each of 17 pre-specified groups within each region across England and Wales. This method is outlined by Census to census matching strategy 2021. This census to census linkage exercise, enabled the estimated proportions of duplication across regions and groups to be estimated with high precision. The estimated proportions of duplication found within each group in each region were then used to calibrate the estimated probabilities of correct enumeration calculated by the model, to produce the final correct enumeration probabilities for each census record. Further information is available in The Proposed Duplication Calibration Method for the 2021 Census of England and Wales.

The estimated level of overcoverage in the 2021 Census of England and Wales was 0.96%, compared to the 2011 Census of England and Wales where the estimated level of overcoverage was 0.6%.

Figure 2a: Age-sex correct enumeration probabilities (female)

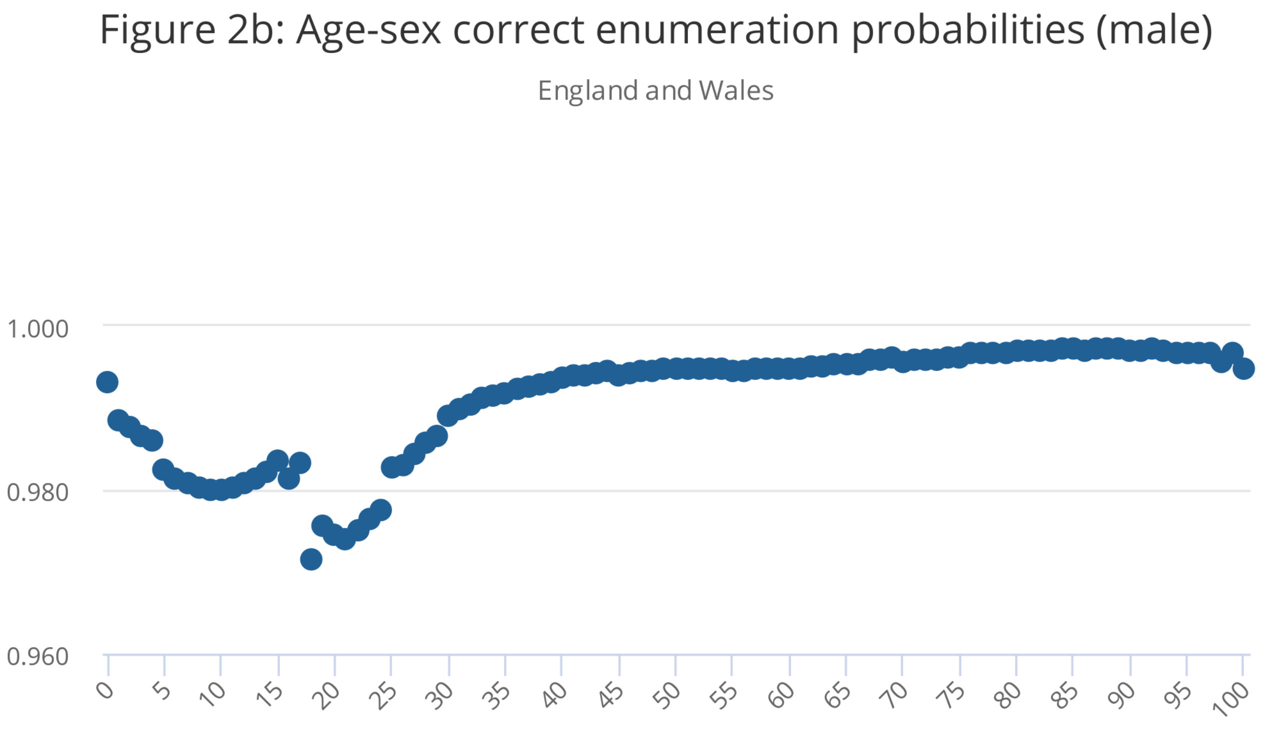
England and Wales



Source: Office for National Statistics, Census 2021

Figure 2b: Age-sex correct enumeration probabilities (male)

England and Wales



Source: Office for National Statistics, Census 2021

5 . Bias adjustments

Based on the experience of the previous censuses, some of the assumptions needed to produce the coverage adjusted population estimated with ignorable levels of bias may not be met in practice. Therefore, the development of methods to adjust for certain biases is also a part of the coverage estimation. In addition, some ad-hoc adjustments may be implemented based on the quality assurance results and availability of the data.

Producing coverage error corrected population size estimates using the Census Coverage Survey (CSS) and census data requires independence between these two data sources. Independence means that for every member of the target population a chance of responding to the coverage survey does not depend on the member being census respondent or non-respondent. In practice, such independence is not achievable and a dependence bias adjustment may be needed. In general, non-responders to census may be less likely to respond to the coverage survey, which would bias the estimates downwards - leading to estimates that are too low. This is the type of bias for which correction was planned and prepared for in advance.

To do this, an alternative estimate was needed. Similar to previous censuses, an Alternative Household Estimate was calculated Alternative Household Estimate 2021. However, since the coverage estimation in 2001 and 2011 Censuses used dual system, ratio, and synthetic approaches, while the coverage estimation in 2021 Census used the mixed-effects logistic regression approach. The way adjustment was applied was very different this time aroundas outlined in Adjusting for the dependence bias in the Census 2021 coverage estimation.

There are two main challenges when using the Alternative Household Estimate to correct for the dependence bias. First, the alternative estimates are available at quite high level of aggregation defined by local authority by hard-to-count index by accommodation type. The second challenge is that reliable alternative estimates are available for the household population only, whereas a 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 the 2021 Census. The approach chosen was the direct adjustment method with reweighting (apportionment) based on the initial undercoverage probabilities.

In 2011, all local authorities were adjusted for dependence bias. However, in the 2021 only five were adjusted.

Another adjustment made was a single year of age adjustment for those aged zero to three years. Based on quality assurance, it was decided to adjust for these age groups across all local authorities in England and Wales using administrative data. In addition, those aged 4 to 15 years were adjusted in Wales and North East based on the School Census.

6 . Other quality assurance adjustments

Several other adjustments were made. There were 15 local authorities in undercoverage estimation were the random effect was forced to be 0 and only the fixed effects part of the model was used. These was due to the fact that the Coverage survey in those areas was not of sufficient quality to reliably support the mixed effects logistic approach, while switching to the logistic regression for the entire country would have had a negative effect for many other local authorities.

The estimated person coverage probabilities in several local authorities were constrained to the household coverage probabilities at the local authority by hard-to-count by accommodation type level. The reason was not directly related to estimation. In this case, the combination of person and household level estimates meant that adjustment process might have experience difficulties. After a careful consideration and assessing the impact, the decision to constrain the probabilities was made.

7 . Variance estimation

Variance estimation measures the variability of the estimates for the key domains of interest like person / household local authority total, local authority by age-sex group total, local authority be tenure. This is outlined in Variance Estimation for 2021 Census Population Estimates. Similarly, to the 2011 Census, the bootstrap method is used. However, unlike the previous census, the bias corrected percentile method was used to produce confidence intervals. This allowed reflecting non-symmetric distribution of the coverage error corrected estimates.

8 . Related links

Item editing and imputation process for Census 2021, England and Wales
Methodology | Released 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.

Model selection for coverage estimation for Census 2021 in England and Wales
Methodology | Released 8 November 2022
The model selection process and chosen models for coverage estimation of Census 2021 in England and Wales.

9 . Cite this methodology

Office for National Statistics (ONS), released 9 November 2022, ONS website, methodology article, Coverage estimation for Census 2021 in England and Wales