Quality assurance of population estimation methods: Census 2021

How we checked and adjusted the estimates of the population for Census 2021 to improve their plausibility and consistency with other sources.

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

Release date:
19 December 2022

Next release:
To be announced

Table of contents

1 . Main points

- Population estimation methods for Census 2021 have several underlying assumptions, which might not always be met in practice, despite these methods being made more sophisticated than ever before by building on 2011 practices.
- It is standard practice to have a series of quality assurance processes to independently check the estimates.
- A thorough quality assurance process led to some adjustments to the census estimates to improve their plausibility and consistency with other sources.
- These adjustments changed the England and Wales total population estimate by 0.08%, though some local authority districts were affected much more than others.

2 . Background

Population estimation for Census 2021 was carried out for people in households and communal establishments. The assurance and adjustments described here relate to the household population only, which represents over 98% of the total population estimate.

Our methods worked well for nearly every area and population group but where our detailed quality assurance identified issues, we used a range of methods to ensure the final estimates were as reliable as possible.

Census 2021 posed several challenges in quality assurance. Even where census estimates are less certain, they are often still better than any alternative as they are part of a large, high-quality exercise. All other sources have either different coverage definitions, partial coverage, or both, which is why we need a census.

For example, the school census in England has good coverage and quality, but only includes school-age children, does not cover private schools or home-schooled children, and may not record a home address consistent with the census definition.

Similarly, the distribution of people by age and sex in each local authority (LA) can be very different to other local authorities and should not necessarily be expected to be consistent. The amount of change since 2011 can also be very difficult to predict. Because all other sources also inevitably have quality issues, and the census is generally high-quality, where there are differences, it can be difficult to know whether it is an anomaly in the census estimates or in the alternative source.

This means checking and potentially adjusting the census estimates is a significant challenge. There were some pre-planned adjustment processes described in Section 3: Adjustments, but not all of these were used in their planned form. This paper describes the adjustments we made, and the logic behind them.

The following adjustments were applied:

- local authority estimates with very high variance were either adjusted up to the Alternative Household Estimate (AHE) where they were lower, or reduced by removing the random effect where they were higher than the AHE
- local authorities where estimation implied that the imputed number of missing households would be greater than the imputed number of persons were constrained to avoid large differences
- the number of very young children was constrained in line with the Personal Demographics Service (PDS), and the number of children aged 3 to 15 years was increased in Wales and the North East because of evidence from the school census

3 . Adjustments

General approach

From experience from prior censuses, we had some planned contingencies for likely issues.

These predicted issues included:

- the need for an overall adjustment as a result of residual bias (PDF, 613 KB): similar to the adjustment needed in 2011 (PDF, 1,300 KB)
- number of students: we were concerned about the possible impact of coronavirus (COVID-19) on our ability to capture the number of students accurately (PDF, 315 KB)
- changes in response patterns because of the changed mode of interview: Census 2021 moved to online-first data collection and we wanted to ensure that this change in mode did not lead to unexpected patterns or results
- response rates for small children: these are always difficult as there is a tendency for small children to be missed from census collection

In practice, there was no evidence that most of these were significant issues in estimating Census 2021, and the only area we made an adjustment for in Estimation was for small children (as described later in this section).

However, while we checked for these issues when evaluating the estimates, our approach was to be open minded and led by the data for where issues may occur. As such, we had also prepared tools to allow us to investigate the estimates and check them for issues more generally.

This quality assurance process involved the estimation team, but also drew on a mix of expertise from across the Office for National Statistics (ONS) to ensure the adjustments were robust, proportionate and based on evidence. We also used expertise from outside ONS, as described in our How we assured the quality of Census 2021 estimates methodology.

Adjustments to areas with high variance

While the modelling approach we used in 2021 allowed us to use information from across England and Wales in estimating each area, some areas will inevitably perform less well than others. These will typically be areas where the census or Census Coverage Survey (CCS) response was low, or where the CCS indicated very different response rates for different parts of the same local authority (LA). Indeed, investigation into the areas with the very highest indicative estimates of variance gave us confidence that the uncertainty in the estimates was caused by data collection issues in CCS rather than by real variability in the census response rates. Additionally, in these areas the estimates often differed most substantially from our alternative sources, also indicating that there was a collection issue.

However, there was no obvious way to set a cut-off for what counted as an implausible level of variance to treat as an outlier. We began with the approach of using something similar to the "Grubbs Test". We calculated the third quartile, interquartile range (IQR), and then used the third quartile plus 1.5 times the IQR as an indicative upper limit for an acceptable level of variance.

This identified 19 LAs, of which 15 had estimates lower than the Alternative Household Estimate (AHE) and four that were higher. Read more in our Using the alternative household estimate (AHE) with Census 2021 methodology.

Plotting the data confirmed that this did appear to be a sensible breakpoint. We then examined the data around this breakpoint on a LA by LA basis to further confirm that this cut-off was well placed. This review used all the data available for quality assurance at the time (from the AHE, 2011 Census, Council Tax, school census, and so on) as well as looking at internal consistency.

Having done these checks, we decided that the best way to treat the outliers was to constrain to our best alternative source, the AHE. The AHE is designed as a way of measuring the number of households without dual-system estimation. While it does not capture individual people, it is thought to be a strong estimate of the number of residential households. This is validated by its correspondence to the modelled estimates in most areas.

The AHE had already been planned as a contingency for census estimation in the event of dependency bias. Dependency bias happens where the responses to the census and CCS are correlated, for example, if people who do not respond to the census are also more likely to not respond to CCS. In this situation, dual-system estimation will be biased low. As such, our contingency had been largely based on the prospect of adjusting the census estimates upwards in those areas where we had evidence of dependency bias. For example, in the 2011 Census, we used an overall adjustment to correct for substantial dependence bias in the responses of young men. As such, while we had planned to constrain our modelled estimates to the AHE in some areas, it had been with a view to increasing the estimates.

However, it was our judgement that the Census 2021 estimates were not, in general, subject to significant dependence bias. This was probably because of the high response to the census itself. Further, the data collection for the census and CCS were more different to each other in 2021 than in previous censuses. The Census 2021 was primarily online, but CCS remained a face-to-face interview.

Using the same AHE alignment for those areas where variance was high, we found that if the AHE was lower than the modelled estimate, our method of application gave some anomalies in the resulting person estimates. Instead, we used an alternative approach where we removed the random effects from the undercoverage model in these areas.

Further information on the random effects can be seen in our Coverage estimation for Census 2021 in England and Wales methodology, but briefly these are the component of the model that tailors it to each local authority specifically. While we control explicitly for many factors in the model, where possible we use random effects to allow for response in each LA to differ somewhat from the others, even after these controls. These are especially important in the most unusual parts of England and Wales, as it allows for variation that cannot be directly captured by the model, but in less heterogeneous LAs they will have less use.

In those LAs where the variance was highest, this would indicate that the data collected for Census 2021 and CCS were sparser and/or less consistent than elsewhere. Much of the variance in these cases would be captured within these local effects, and so removing the random effects and estimating from the broader model alone reduces the variance substantially.

Note that response estimates in these areas are still based on the model, and so still controlled for all the variables within that model.

Having removed the random effects, the results were again checked. In all cases the plausibility of the refined estimates was improved. The household estimates were similar to the AHE, but with improved person estimates than using that method.

It is important to note that while this process was used to identify potential issues, a large amount of less formulaic quality assurance was also undertaken. Results for LAs were reviewed for internal consistency and compared with the other data available. This assurance led to the other adjustments made and was also used to validate the application of the adjustments we have described.

In particular, while Middlesbrough had not been identified as a high-variance outlier in this process, it was one of the remaining areas most different to the AHE. Further investigation suggested that the modelled results were anomalous when compared with our other sources, suggesting a data collection issue. We implemented the same approach as for the areas with high variance, removing the random effect from the model.

Adjustment in areas where imputed persons was less than imputed households

Having made the first set of adjustments, we reviewed the data again, including consideration of census processes further on in the processing pipeline.

We found that in some areas the number of estimated missing households was higher than the number of estimated missing people. In principle, this is not a defect for coverage estimation because the counted population includes person-level overcoverage, and there is unavoidable inherent uncertainty in both the person and household level estimates. For more information on overcoverage, please see our Coverage estimation for Census 2021 in England and Wales methodology.

However, record imputation is the process following estimation. This process imputes households and persons within them so that the census database is consistent with the estimated population figures. This process cannot impute empty households, nor could we remove responding persons from existing households. As such, this process generally expects the number of imputed persons to be greater than the number of imputed households. We decided to constrain areas where this requirement was especially difficult so that adjustment would work smoothly. More information is available in our Coverage Adjustment for Census 2021 in England and Wales methodology.

In general, our estimates of households are easier to validate than our estimates of individuals, because we have the AHE and more administrative data at the household level. Therefore, we decided to increase the numbers of imputed persons in areas where the imputed number of households was at least 50 more than the number of imputed persons (rather than reducing the number of imputed households). In these areas, we increased the number of imputed persons to be equal to the number of imputed households. This affected 10 local authorities.

Adjustments to estimates of children

As mentioned, the census estimates were compared with available administrative data. This included data on registered births, and both the school censuses from England and Wales.

A review of the estimates of children aged zero to two years showed that they seemed low. This was both in comparison with the adjacent age groups, and in comparison with the Personal Demographics Service (PDS).

While migration (inward or outward) and deaths can cause differences between the PDS and census, for very young children these should be small. Further, it is a known issue for census collection that very young children are sometimes missed. This is a pattern we see in most countries taking censuses around the world. The mechanism for this is not entirely clear but could be as simple as parents forgetting to complete a form for their child.

We decided to constrain the estimates of children aged zero to two years to be at least those recorded in the PDS. This had an impact in almost all LAs, though often in very small numbers.

As a further part of the quality assurance process, it was identified that the estimates for children aged 5 to 15 years were somewhat lower than the number suggested by the school census in Wales. This comparison is more difficult than for the estimate for children aged zero to two years. The school a child attends is not necessarily in the same LA as where the child lives, private schools do not complete the school census and home-schooled children are also not included.

However, while these factors are present, using the best available estimates for each of these indicated that the estimates for children in Wales appeared low, and so we calculated an uplift factor in line with these figures.

Having determined that this was the case, we reviewed similar evidence for LAs in England. In general, the evidence here was less strong and less consistent. However, we did find that in the North East region, there appeared to be a similar pattern to that of Wales, with a lower estimated population than suggested by the schools census. As such, we decided to make a similar change with the North East estimates. Because the data in Wales were more complete, we used the same uplift factor.

As the changes outlined would result in an uplift to children aged zero to two years and those aged 5 to 15 years, this could lead to an inconsistency with those aged 3 to 4 years. We uplifted this age group by the same factor as those aged 5 to 15 years to ensure consistency and a smooth demographic pattern in those regions.

Notes on the attached adjustments table

A processing error meant that the adjustments described were not correctly applied for 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 setting the random effect in the model to zero. 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.

The area subject to the largest adjustment was Gwynedd, which we therefore examined closely. There were special difficulties with the CCS in some postcodes in this LA, which led to the random effect being unstable. As such, removal of the random effect made a large improvement, producing estimates that were much more plausible compared with our comparator data.

Because of the complex processes involved in introducing constraints to the model, some small changes may apply even where a LA was not targeted for a change.

More detail can be found in our collated adjustments for Census 2021 spreadsheet.

4 . Related links

Coverage adjustment for Census 2021 in England and Wales
Methodology | Released 19 December 2022
Methodology for the coverage adjustment of Census 2021 in England and Wales.

Using the alternative household estimate (AHE) with Census 2021
Methodology | Released 9 December 2022
Methodology for the alternative household estimate and how we used it to validate final dual system estimation (DSE) estimates.

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.

Model selection for coverage estimation for Census 2021 in England and Wales
Methodology | Last revised 9 November 2022
The model selection process and chosen models for coverage estimation of Census 2021 in England and Wales.

How we assured the quality of Census 2021 estimates
Methodology | Released 7 November 2022
Methodology for the validation of Census 2021 population estimates for England and Wales, including the assurance of processes, assessment of estimates, and involvement of local authorities.

5 . Cite this methodology

Office for National Statistics (ONS), released 19 December 2022, ONS website, methodology, Quality assurance of population estimation methods: Census 2021