Variance Estimation for 2021 Census Population Estimates

Gareth Powell and Viktor Racinskij, ONS

Summary

This paper outlines the methods considered for estimating confidence intervals for Census 2021, proposing a bootstrap approach to calculating variance and empirical calculation of non-symmetrical confidence intervals. We ask the panel to note this recommendation.

A key stage of the 2021 Census is to estimate the population size by key groups, accounting for under- and over-coverage. This is done using dual-system estimation (DSE), based on the Census Coverage Survey (CCS), and the approach has been outlined in previous MARP papers (see EAP103, EAP105, EAP155 but also EAP112, EAP127, EAP128, EAP161). As key context and quality information for these estimates, there is a requirement to produce confidence intervals which inform users about uncertainty. These help users to understand the precision of the estimates and help them to be used correctly.

Based on work commissioned by ONS by Prof James Brown (see annex) and further work and testing by ONS, three approaches to variance estimation for Estimation in the 2021 Census have been considered: Random Groups, Jackknife and Bootstrap. Of these, and following Prof Brown's recommendation, this paper proposes use of the Bootstrap method. This is the most practical and flexible method and builds on the experience of its successful application in the 2011 Census. This choice is based on that previous experience as well as previous experience, rather than simulation

studies, due to the difficulties in developing working methods for the other two approaches.

We also describe here the method for our construction of confidence intervals, which can now be estimated empirically.

Background

The CCS has a two-stage approach in which output areas (OAs) are the primary sampling units (PSUs) and postcodes are the secondary (final) sampling units (SSUs). See EAP127 for detail. The sample is stratified by Local Authority (LA) and Hard-to-Count (HtC) level. The sample is distributed optimally at hard-to-count level only (five strata), and then allocated to local authorities within each hard-to-count index proportional to the size of each Local Authority. A minimum of 1 PSU per stratum and a maximum of 120 per local authority[1] are also used to ensure spread and avoid overrepresentation.

CCS returns are matched to Census, allowing us to use DSE techniques to account for under- and over-coverage. In 2011 (Baillie et al, 2010), this was done by using stratified DSE, within each LA/HtC strata, while in 2021 we will be using logistic regression modelling to estimate response propensities.

The key outputs of this modelling will be population estimates by LA, age-sex group and similar key variables. These will then feed in as constraints to the Adjustment household imputation process (EAP122). As estimates, these values are uncertain,

[1] This constraint was relaxed for Birmingham and Leeds

and the degree of uncertainty should be measured as a key context for the results, as well as ensuring that users understand how precise the estimates are. Finally, one of the Census success criteria is that it should be nationally accurate as measured by a confidence interval of +/-0.2%.

In both 2001 and 2011, the methods used to produce confidence intervals resulted in symmetrical estimates (95% confidence intervals for the 2001 Census result and Census General Report for England and Wales (2011), Chapter 8). However, there will be substantial skew in practice, as undercoverage is a greater issue than overcoverage.

Random Groups

To implement a Random Groups approach (Wolter, K. M (2007), Chapter 2), we would randomly partition the sample into x non-overlapping groups, carry out estimation on each of the random groups and estimate the variance based on variation between the groups.

However, the random groups approach requires each random group to be formed according to the design of the parent (original) sample. For a stratified sample (which is the case for the Census Coverage Survey) this means that each random group is itself a stratified sample. In addition, when cluster sampling is used (which also the case in CCS), random groups are formed by selecting primary sampling units (Wolter 2007, pp. 32 – 33). The design and allocation of the 2021 Census coverage survey is such that each stratum (local authority by hard-to-count group) has at least two primary sampling units (output areas). This means that only two random groups can be formed while maintaining the parent design.

It is possible that some compromise between the number of groups and the original design could be made. For instance, in model selection we are using 5-fold cross validation. However, splitting the data into the folds must preserve the design of the parent sample as much as possible and even splitting into 5 groups results in some of the strata or even local authorities being collapsed together. Variance estimation using the random groups approach may require forming somewhere between 10 to 30 random groups and the amount of collapsing and resulting departure from the design of the parent sample is likely to lead to unstable estimates of variance, with potential biases and lack of good coverage.

Another potential problem with the random groups approach is that there may be model convergence issues when the number of random groups is large. The variance estimation within each random group would use a fixed coverage estimation model specification. As the number of random groups increases, the chance that the selected model will not converge for some of the random groups also increases. The Bootstrap approach is more reliable in this respect.

A repeated balanced half-sample method could be used to overcome some of the above issues. However, we did not have time to investigate this option fully.

Jackknife

This method is discussed in Chapter 4 of Wolter (2007). In this approach, we would first partition the sample into *k* groups of *m* observations (*k* may be equal to the sample size). We then leave out one group in turn, estimate the population size on the remaining data and estimate the variance based on the estimates across all the

groups. This approach was used in the 2001 Census, leaving out one primary sampling unit within each estimation / design stratum. The jackknife was well justified in the context of 2001 Census since the sample design and estimation strata coincided. However, this approach is not as straightforward when the estimator pools across strata.

Use of the jackknife in for the coverage estimation with estimation strata pooling across the design strata has not been fully investigated and is not well developed in the literature for census coverage applications. However, it is likely to be problematic given the 2021 Census estimation strategy. For example, it would require an impractical amount of replication as we are pooling across large parts of the coverage survey sample for estimation components in 2021.

Bootstrap

Used in the 2011 Census, this approach mimics the sampling distribution of the parameter of interest through generation of repeated pseudo-samples with replacements from the actual sample (Efron and Tibshirani, 1993).

The difficulty with the bootstrap is its correct application within finite population sampling and this is covered in detail in Chapter 5 of Wolter (2007). The issue is the correct creation of pseudo-samples that reflect sampling without replacement from a finite population as well as the multi-stage stratified design in the CCS. If this is not addressed, then bootstrap pseudo-samples from the pseudo-population will not mimic the realised sample from the actual population (Efron and Tibshirani, 1993). We deal with the multi-stage structure as we did for random groups and jackknife; we build pseudo-samples by re-sampling the PSUs but not re-sampling within the

PSU[2]. We treat the selected PSU as a 'block' and then effectively treat blocks as interchangeable.

Embedding all the other sample structure, such as stratification and sampling without replacement, is also important. Two approaches are possible. One approach, which goes back to Gross (1980), is to build a pseudo-population by first copying the selected PSUs $\frac{1}{\pi}$ times where π is the selection probability for a sampled PSU. This creates a pseudo-finite population, and we can now simply apply our chosen design repeatedly to create pseudo-samples that have the correct stratification and PSU structure. As the pseudo-samples embed the correct design strata, we can implement estimation that crosses those strata and the bootstrap variance will reflect this. For simple random sampling without replacement, Wolter (2007) shows this approach to be biased (downwards).

An alternative, put forward in Wolter (2007), is to sample the PSUs with replacement from each design strata selecting $(n_h - 1)$ rather than n_h PSUs from each stratum hThis gives a variance estimator that corrects for the bias discussed previously but omits the finite population correction factor. However, even in the case of the large coverage survey, this is likely to have a negligible effect.

We will use this approach as it is less error prone and simple implementation-wise and the fact that omitting the finite population correction factor will have a little effect on the variance estimates. Estimating the variance of the parameter of interest via bootstrap typically requires between 50 and 200 replicates (Efron and Tibshirani, 1993).

[2] Early exploration for 2011 re-sampled postcodes within PSUs and demonstrated empirically that this over-estimated the variance and is an unnecessary step.

The bootstrap method, as described, is therefore the approach we will take to the estimation of variance.

Estimating Confidence Intervals

Once the variance estimates are obtained using the bootstrap method as described, we can utilise the large sample properties to construct the intervals. This how the intervals were constructed back in 2001 (with Jackknife) and 2011 (with Bootstrap). The drawback of this approach that the resulting confidence intervals are symmetric. However, the ‘true’ intervals are substantially skewed.

While allowing skewed intervals is a more accurate reflection of the truth, as well as not being possible for some approaches it is also perhaps less well understood by users, who may be used to, and expect, symmetrical intervals. This is an issue of communication however and we intend to produce the best estimates possible and then explain them.

Estimates of variance will be available for the key estimates - household totals, person totals, persons by age-sex and persons by ethnicity. Other variance estimates will be considered and discussed with the outputs team.

If we increase the number of bootstrap replicates to more than 1000, it becomes possible to estimate the limits of the confidence interval empirically from the generated bootstrap distribution. This allows us to produce non-symmetric intervals. However, with 2000 replicates, when constructing a 95% confidence interval the limits are still defined by just the 50 smallest and largest replicates. Therefore, the end-points can be unstable. So, to obtain the stable empirical endpoints, a large

number of replications are needed. Optimization and parallelization work allows us to run a large number of (6,000 – 10,000 subject to model complexity) in a reasonable time. As such, this time we aim to produce non-symmetric intervals empirically. Estimates of variance will be produced immediately after population estimation, and so confidence intervals should be available for the first release of Census 2021 estimates.

Conclusions

Given the success of the bootstrap approach in 2011, this should be the preferred approach for 2021 Census. It offers the most flexibility in being able to estimate the variance of an estimator that combines under-count, over-count, and dependence adjustments, which are each estimated from different (but typically overlapping) samples. Empirical confidence intervals are now computationally possible, allowing non-symmetrical confidence intervals and improving on the 2011 methods.

References

EAP103 – Census Coverage Survey Design Strategy

EAP105 – Coverage Estimation Strategy for the 2021 Census of England and Wales

EAP112 – Over-coverage estimation strategy for the 2021 Census of England and Wales

EAP122 – The 2021 Census Coverage Adjustment

EAP127 – CCS 2021 allocation strategy

EAP128 – Informative sampling in coverage estimation of Census 2021

EAP155 – Coverage Estimation for Small Communal Establishments

EAP161 – Residual bias adjustment

Baillie, Brown, Taylor and Abbott, 2010. Available at: https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howwetookthe2011census/howweprocessedtheinformation/coverageassessmentandadjustmentprocesses/imagesvarianceestimationv1tcm774049_tcm77-224787(2).pdf

Census General Report for England and Wales, Chapter 8. Available at: https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howdidwedoin2011/2011censusgeneralreport/2011censusgeneralreportforenglandandwaleschapter8_tcm77-384975.pdf

Efron, B. & Tibshirani. R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall: New York.

Wolter, K. M (2007). Introduction to Variance Estimation. 2nd Edition. New York: Springer.

Annex

Overview of Approaches to Variance Estimation for Estimated Census Adjustment

Prof James J Brown (University of Technology Sydney)

This paper briefly reviews the main replication approaches for variance estimation, drawing on the key texts of Wolter (2007), and Efron and Tibshirani (1993). The recommendation is to pursue a bootstrap approach for variance estimation but further work is required to confirm that empirical confidence intervals, with stable end-points, is feasible. Variance estimation for the imputed database is discussed and a simplified quality measure, which does not require replication, is developed. However, its appropriateness will require testing through some replication applied to the 2011 Census database.

1) Introduction

A key component of any set of survey estimates is the accompany set of estimated standard errors and associated confidence intervals to communicate the uncertainty in the estimates due to sampling error. With estimates of the census population adjusted for coverage, these are crucial as they allow us to:

1. demonstrate clearly that the adjustments have reduced the overall error compared to the census,
2. show that the final adjusted estimates are in-line (consistent) with other sources of data such as administrative counts and historical rolled forward estimates.

With 1) we can demonstrate that the sampling error in our adjusted estimate is smaller than the implied bias in the unadjusted census. If our adjusted estimate is $\hat{T}$ with an associated estimated standard error $\hat{\sigma}$, and the (unadjusted) census count is C; we can demonstrate that the empirical mean-square-error of the adjusted estimate, $\widehat{MSE}(\hat{T}) = \hat{\sigma}^2$ is less than $\widehat{MSE}(C) = (\hat{T} - C)^2$. With 2) we can use our confidence intervals on the adjusted estimates to help quantify if differences between the adjusted estimates and other sources are due to sampling error or some other underlying systematic error (in either the adjusted estimates or the comparison sources).

In 2001, estimation of standard errors used a jackknife approach (Brown, 2000); while in 2011 a bootstrap approach was implemented (Brown *et al*., 2019). The move to the bootstrap was based on work reported in Baillie *et al*. (2011). For both censuses, the calculation of confidence intervals was based on the assumption that the sampling distribution of $\hat{T}$ would be approximately Normal, although Baillie *et al*. explored the possibility of estimating empirical confidence intervals directly

from the bootstrap distribution. Empirical confidence intervals are not constrained to be symmetric. This can be advantageous when working with estimators that have lower bounds, and the estimated value is close to that point. You do not create intervals that go below the lower bound and correctly assign wider uncertainty above the estimate. With the adjusted estimate $\hat{T}$, the census count C is a soft lower-bound, so for adjusted estimates close to the census a symmetric confidence interval can go below the census count while not correctly recognising the upwards uncertainty. It is a soft lower-bound because over-count adjustments can result in an adjusted estimate below the census but in the UK context that has not occurred.

Section 2 gives a brief overview of variance estimation approaches, with a focus on jackknife and bootstrap; and discusses the construction of confidence intervals. Section 3 considers the application of a bootstrap approach in 2021 and Section 4 finishes with some discussion of variance estimation for the final database.

2) Approaches to Variance Estimation

The aim of this section is not a thorough review of all approaches (see Wolter, 2007 for such a review) but to give an overview. Broadly speaking, we can either estimate variances via analytical approaches or through replication approaches. With complex estimators, such as those applied to census adjustment, a direct analytical approach is not usually feasible and the approximations via Taylor Series linearization are complex as well. Replication methods essentially approximate the sampling distribution of an estimator by repeated application of the estimation strategy to a set of pseudo-samples. Differences in replication methods relate to that creation of the pseudo-samples.

2.1) Random Groups

This approach is covered in detail in Chapter 2 of Wolter (2007). Ideally, we have independent random groups but this involves selecting a set of (small) independent samples rather than one single sample. In reality, dependent groups are formed by partitioning the overall sample once it has been selected. In the context of the 2001 Census, Brown (2000) explored this approach using single primary sampling units (PSUs) to create the random groups, with the between cluster variation as an estimate of variance. In others words, if our estimate $\hat{T}$ for an estimation area[3] (EA) is based on a sample of $k = 1, \dots, K$ PSUs and $\hat{T}_k$ is the estimator applied to a single PSU, then a simple variance estimator is given by

$$\frac{1}{K(K-1)} \sum_k \left(\hat{T}_k - \hat{T}\right)^2.$$

[3] In both 2001 and 2011, the Census Coverage Survey design and estimation is based on around 100 estimation areas for England and Wales, with estimation essentially being implemented independently within each area (Brown *et al*., 2011; Brown *et al*., 2019).

In the application to the 2001 Census approach, simulations in Brown (2000) showed this approach over-estimated the variance while not giving robust coverage for confidence intervals. In part, this was due to the non-linear structure of the estimator for 2001 that had robust adjustments applied to the ratio model driving estimation. The robust adjustments were dropped in 2011 and therefore we might expect better performance from this (simple) replication approach. However, the estimation strategy has increased its complexity in other ways, such as components pooling across the EAs. Therefore, random groups would need to be formed by partitioning the PSUs for the entire national sample, while respecting the design structure within each EA[4], rather than forming groups and carrying-out variance estimation independently within each EA. Once these national random groups are formed, the national estimation strategy is applied to each overall sample group rather than independent application within each EA.

2.2) Jackknife

This approach is covered in detail in Chapter 4 of Wolter (2007). As with random groups, we split the sample up into a set of groups, in the case of multi-stage designs these are constructed from the PSUs respecting stratification. To create the pseudo-samples we just drop one group (at a time) and re-estimate. In the context of the 2001 Census, Brown (2000) explored this approach dropping single primary sampling units (PSUs) to create the random groups. Following a similar notation to Section 2.1, a jackknife variance estimator for $\hat{T}$ has the form

$$\frac{1}{K(K-1)}\sum_k \left(\hat{T}_k - \hat{T}\right)^2 = \frac{(K-1)}{K}\sum_k \left(\hat{T}_{(not\ k)} - \hat{T}\right)^2$$

where $\hat{T}_k = K\hat{T} - (K-1)\hat{T}_{(not\ k)}$, and $\hat{T}_{(not\ k)}$ is the estimator based on all groups apart from k. In this simple form, estimation must be implemented independently within strata and this was the case in 2001. If $\hat{T}$ is an estimator that pools across strata, as was the case in 2011 where estimation strata collapse across design strata and will be in 2021 with high-level estimation models, then the application is not quite so straightforward. Following Wolter (2007), one PSU is dropped from the entire sample to create

$$\sum_h \frac{q_h}{n_h}\sum_{k=1}^{K_h}\left(\hat{T}_{h(not\ k)} - \hat{T}\right)^2 \text{ (see equation 4.5.6; Wolter, 2007)}$$

where $\hat{T}_{h(not\ k)}$ estimates $\hat{T}$ with a single PSU k dropped from stratum h, $q_h = (K_h - 1)(1 - K_h/N_h)$, K_h is the number of PSUs in stratum h, and N_h is the number of PSUs in the population. Such an approach has not been evaluated for UK census adjustment and potentially requires a considerable amount of replication if we are pooling across large parts on the national coverage survey sample for estimation components in 2021.

2.3) Bootstrap

[4] Balance across the full sample structure may be hard to achieve as some strata at the design stage have small numbers of PSUs selected limiting the number of groups that can be formed if those strata are to be correctly accounted even when pooling strata at estimation.

The bootstrap approach was introduced in Efron (1979) and has intuitive appeal as a tool for variance estimation. It mimics the actual sampling distribution of $\hat{T}$ through generating repeated pseudo-samples by sampling with replacement from the achieved sample. Ignoring the issue of how to create those pseudo-samples, if we create $b = 1, \ldots, B$ pseudo-samples, and $\hat{T}_b$ is the estimate based on the b^{th} pseudo-sample, then the bootstrap variance estimator has the form

$$\frac{1}{(B-1)}\sum\nolimits_k \left(\hat{T}_b - \hat{\bar{T}}\right)^2$$

where $\hat{\bar{T}} = \frac{1}{B}\sum_b \hat{T}_b$. This was implemented in 2011 Census to estimate standard errors allowing the integration of components for over-coverage and dependence adjustment, which are not necessarily estimated at the same level of aggregation as the main under-coverage component. Standard errors of estimates of functions of estimates, such as the sex ratio, are easily computed by estimating the quantity within each pseudo-sample b and applying the variance estimator (Wolter, 2007).

The difficulty with the bootstrap is its correct application within finite population sampling and this is covered in detail in Chapter 5 for Wolter (2007). The issue is the correct creation of pseudo-samples that reflect sampling without replacement from a finite population as well as the multi-stage stratified design. If we do not do this then our bootstrap pseudo-samples from the pseudo-population will not mimic our realised sample from the actual population (Efron and Tibshirani, 1993). We deal with the multi-stage structure as we did for random groups and jackknife; we build pseudo-samples by re-sampling the PSUs but not re-sampling within the PSU[5]. We treat the selected PSU as a 'block' and then effectively treat blocks as exchangeable.

Embedding all the other sample structure, such as stratification and sampling without replacement, is also important. One approach, which goes back to Gross (1980), is to build a pseudo-population by first copying the selected PSUs $\frac{1}{\pi}$ times where π is the selection probability for a sampled PSU. This creates a pseudo-finite-population and we can now simply apply our chosen design repeatedly to create pseudo-samples that have the correct stratification and PSU structure. As the pseudo-samples embed the correct design strata, we can implement estimation that crosses those strata and the bootstrap variance will reflect this. For simple random sampling without replacement, Wolter (2007) shows this approach to be biased (downwards) for the variance by a factor $\frac{Nn-N}{Nn-n}$, where N is the population size and n is the sample size. In the context of selecting PSUs, it makes intuitive sense that in those strata where n is small this has the biggest impact because the pseudo-population is built from very few PSUs. With the coverage survey design, this will be an issue with a small number of the design strata, and additional simulation work to test this impact would be desirable as the research in preparation for 2011 worked with a design where it was not necessary to collapse strata at estimation due to small sample sizes.

[5] Early exploration for 2011 re-sampled postcodes within PSUs and demonstrated empirically that this over-estimated the variance and is an unnecessary step.

One final issue is that $\frac{1}{\pi}$ is rarely an integer. A simple solution (Wolter, 2007) is to choose randomly whether to expand a PSU by its integer component or add one. The choice can be made independently with the selection of each bootstrap sample. Alternatively, in 2011 the pseudo-population was built by sampling with replacement within strata from the selected PSUs, which on average mirrors expanding by $\frac{1}{\pi}$.

An alternative, put forward in Wolter (2007), is to sample the PSUs with replacement from each design strata selecting $(n_h - 1)$ rather than n_h PSUs from each stratum h. This gives a variance estimator that corrects for the bias discussed previously but omits the finite population correction factor. However, even in the case of the large coverage survey, this is likely to have a negligible effect.

2.4) Constructing Confidence Intervals

As discussed in Section 1, once we have an estimated standard error we also wish to construct confidence intervals. Typically, we draw on statistical theory to assert that in large samples the property

$$\frac{\hat{T} - T}{\sigma} \sim N(0,1) \rightarrow \frac{\hat{T} - T}{\hat{\sigma}} \sim t_{n-1}$$

holds, where σ^2 is the sampling variance of $\hat{T}$, $\hat{\sigma}^2$ is its estimate via one of the replication methods above, and t_{n-1} is the t-distribution with number of PSUs minus one as the degrees of freedom. An estimated $(1 - \alpha)\%$ confidence interval is then simply

$$\hat{T} \pm t_{n-1,\frac{\alpha}{2}} \times \hat{\sigma},$$

and this was the approach implemented in both 2001 and 2011. Estimating $\hat{\sigma}^2$ via bootstrap typically requires between 50 and 200 replicates (Efron and Tibshirani, 1993). If we increase the number of bootstrap replicates to more than 1000, it becomes possible to estimate the limits of the confidence interval empirically from the generated bootstrap distribution. However, with 2000 replicates, when constructing a 95% confidence interval the limits are still defined by just the 50 smallest and largest replicates. Therefore, the end-points can be unstable.

For 2011 Census, Baillie *et al*. (2011) explored implementing empirical confidence intervals, and found that the so-called BCa approach worked effectively (Efron and Tibshirani, 1993). BCa stands for bias corrected and accelerated. Intuitively, empirical confidence intervals are easy. If we wish to estimate a $(1 - \alpha)\%$ confidence interval, order the B bootstrap replicates, and then choose the $B \times \frac{\alpha}{2}$ and $B \times \left(1 - \frac{\alpha}{2}\right)$ replicates as the end-points. The first issue is that the sampling distribution of $\hat{T}_b$ does not (precisely) centre on $\hat{T}$. The bias correction fixes this component by estimating the shift in the median of the distribution of $\hat{T}_b$ relative to $\hat{T}$. It calculates the shift $\hat{z}_0$ based on inverting the cumulative normal distribution for the estimated probability (proportion) that $\hat{T}_b < \hat{T}$.

Therefore, if exactly half the bootstrap replicates give an estimate for $\hat{T}_b$ that is less than $\hat{T}$, then $\hat{z}_0 = 0$.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{count\left[\hat{T}_b < \hat{T}\right]}{B}\right)$$

The acceleration component relates to the 'rate of change of the standard error of the estimate with respect to the true parameter value' (Efron and Tibshirani, 1993; p186). It essentially adjusts for skewness in the bootstrap distribution to get more stable end-points for the empirical confidence interval. The acceleration parameter a is estimated via a jackknife applied to the original sample data such that

$$\hat{a} = \frac{\sum_k \left(\hat{T}_{(.)} - \hat{T}_{(not\ k)}\right)^3}{6\left\{\sum_k \left(\hat{T}_{(.)} - \hat{T}_{(not\ k)}\right)^2\right\}^{3/2}},$$

where $\hat{T}_{(.)} = \frac{1}{K}\sum_k \hat{T}_{(not\ k)}$ and $\hat{T}_{(not\ k)}$ is as defined in Section 2.2. Now we can estimate the adjustment to the percentile points for a $(1-\alpha)\%$ confidence interval as

$$\alpha_L = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\frac{\alpha}{2}}}{1 - \hat{a}\left(\hat{z}_0 + z_{\frac{\alpha}{2}}\right)}\right) \rightarrow \alpha_L = \Phi\left(z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \text{ if } \hat{z}_0 = 0 \text{ and } \hat{a} = 0$$

$$\alpha_U = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\frac{\alpha}{2}}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\frac{\alpha}{2}}\right)}\right) \rightarrow \alpha_U = \Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} \text{ if } \hat{z}_0 = 0 \text{ and } \hat{a} = 0$$

To select the end-points, order the B bootstrap replicates, and then choose the $B \times \alpha_L$ replicate as the bottom of the confidence interval and the $B \times \alpha_U$ replicate as the top.

A benefit of the BCa confidence intervals is that if we take a (monotone increasing) function of $\hat{T}$ then we can apply the same function to the end-points of the BCa interval (Efron and Tibshirani, 1993). However, we are more interested in calculating quantities that are functions of several estimates. As an example, we might want to get the sex ratio as $\frac{\hat{T}_m}{\hat{T}_f}$. Provided we have kept the bootstrap replicates then the estimated variance would be

$$\frac{1}{(B-1)}\sum_k \left(\frac{\hat{T}_{mb}}{\hat{T}_{fb}} - \frac{\overline{\hat{T}_m}}{\hat{T}_f}\right)^2$$

where $\frac{\overline{\hat{T}_m}}{\hat{T}_f}$ is the average of the sex ratios across the bootstrap replicates. To apply the BCa we would also need to do the jackknife step for $\frac{\hat{T}_{m(not\ k)}}{\hat{T}_{f(not\ k)}}$ to estimate the specific acceleration parameter. If we have also stored all the jackknife estimates then this is also straightforward but is still an additional step. It then gets more complex if $\hat{T}$ is made-up of several components coming from over-lapping (but not identical) parts of the national sample. How we apply the jackknife is not so simple, and this is of course one of the bootstrap advantages. It should also be noted that as discussed in Section 2.2, and Wolter (2007), care is needed when working with stratified PSUs, and the development of $\hat{a}$

does not account for this[6]. The current assumption would be to jackknife PSUs across the whole sample ignoring the stratification structure but this would be computationally expensive.

As pointed-out in Section 1, the attraction of an empirical confidence interval is that it no longer enforces symmetry on the confidence interval; and that can be useful with census coverage as the estimate has a (soft) constraint driven by the actual census count. However, empirical bootstrap is not the only approach. Kabzinska, Smith, and Berger (2017) developed an empirical likelihood confidence interval, which also does not enforce symmetry. The approach does not require the bootstrapping but is not as flexible in terms of incorporating different components into the estimation of $\hat{T}$, where those components depend on different units from the national sample. However, we also note that applying BCa in such situations is also not straightforward.

2.5) Conclusion

Given the success of the bootstrap approach in 2011, this should be the preferred approach for 2021 Census. It offers the most flexibility in being able to estimate the variance of an estimator that combines under-count, over-count, and dependence adjustments, which are each estimated from different (but typically overlapping) samples. Re-visiting empirical confidence intervals is also desirable as simulation work in 2011 (Baillie *et al*., 2011) showed these to offer excellent coverage properties. However, the BCa approach may only be practical for a set of pre-specified estimates, where the jackknife structure is also pre-specified.

The decision to pursue bootstrap does not lock ONS into any specific estimation strategy. In fact, provided the pseudo-samples can be generated 'correctly', we can then adopt whatever estimation strategy we wish, as long as there is sufficient computing resource to replicate it across the bootstrap pseudo-samples[7]. Note that simply estimating $\hat{\sigma}$ requires considerably less replication than for empirical confidence intervals.

3) Implementing the Bootstrap

Given the discussion in Section 2, we consider a possible implementation strategy. The sample of PSUs for the coverage survey will be drawn ahead of time and known. Therefore, it is possible to select and store all the pseudo-samples of PSUs. Once the data for the PSUs is processed, it can be attached to pseudo-samples and then generating the bootstrap replicates just requires multiple runs

[6] We note that the simulation work for 2011 (Baillie *et al*. (2011)) found the BCa to have excellent coverage properties when applied with bootstrapping and a simple stratification structure.

[7] The current default for 2021 estimation builds large logistic regression models as the basis for the under-count component, this may put a practical constraint on the number of bootstrap replicates, which would just result in t-distribution confidence intervals rather than empirical confidence intervals.

of the estimation system. In a similar way, the PSUs for the jackknife replicates for the required acceleration parameters can also be pre-specified and stored ready to be used.

The approach to building the pseudo-samples taken in 2011 first (re-)built a pseudo-population by re-sampling the selected OAs with replacement. Bootstrap replicates were then drawn from that pseudo-population. Simulation work in Baillie et al. (2011) showed this performed well but the design-stratification in the simulations was not as detailed as in the actual 2011 design. In other words, we are less sure of the impact of the smaller strata on the bias discussed at the start of Section 2.3. It would be prudent to repeat the simulation work for an EA (or set of EAs) using the 2011 design that stratified by LA but ignored the LA stratification in estimation; and compare to the alternative approach suggest by Wolter (2007) of just re-sampling from each stratum with a bootstrap sample size of $(n_h - 1)$.

A comprehensive simulation study would also allow testing of the logistic regression approach to under-coverage estimation, which fits across a large number of the design strata. At this stage, it is likely unrealistic to integrate over-coverage and dependence components as well in the simulation. Further exploration of BCa empirical confidence intervals would also be possible, and this would allow testing of implementing the jackknife across design strata.

4) Variance Estimation and the Adjusted Database

Post the 2011 Census, Sexton and Brown (unpublished work) explored the variability of the adjusted database by running the imputation system through on a small number of the bootstrap replicates. This draws on the concept of multiple imputation introduced in Rubin (1987). However, there are two issues. One, replicating the full imputation system for even a small number of replicates will require considerable computing resource. Two, it is not clear how we would report or integrate multiple databases into the output system.

4.1) Full Replication

One approach would be to replicate the full imputation system on say the first 11 bootstrap replicates, and then for each potential tabulation report the range of the middle six totals. (Indicates the inter-quartile range for the total on a table.) However, achieving this would still be an un-known for 2021. More full replications seems unrealistic given the computing resources involved with a single run of the full imputation system in 2011.

4.2) Partial Replication

More plausible would be to recognise that the database is created from integer weights, which are themselves the output of a logistic model. In 2011, the imputation system built on the 2001 approach developed in Steele, Brown, and Chambers (2002). It retained the two stages of imputing missed households and missed individuals into counted households, although the ordering was reversed relative to 2001. The working proposal for 2021 is to impute whole households only, and to use constrained optimisation to create integer weights for counted households to meet the estimated control totals. Testing shows that, before placement of imputed households, this approach is computationally efficient and considerably faster than 2011.

Efron and Tibshirani (1993) argue that, with just 50 bootstrap replicates, standard errors are estimated effectively. Given a set of integer weights w_{bi} for units i in replicate b, we would estimate the variance for the total of some (low) level of output A as

$$\frac{1}{(B-1)}\sum_k \left(\hat{T}_b - \hat{\bar{T}}\right)^2$$

where $\hat{T}_b = \sum_{i\in A} w_{bi}$ is the sum of the weights for that area, and $\hat{\bar{T}}$ is the mean across bootstrap replicates. This still requires considerable replication.

4.3) Weighting Approximations

If we consider an estimate of total for area A to be approximately $\hat{T} = \sum_{i\in A_C} w_i = \sum_{i\in A_C} \frac{X_i}{\hat{\pi}_i}$, where $X_i = 1$ for the A_C census responding units and 0 for those units missed in area A, then we can make progress on an approximate variance of $\hat{T}$ as

$$V\left[\hat{T}\right] = E\left[V\left[\hat{T}|\hat{\pi}_i\right]\right] + V\left[E\left[\hat{T}|\hat{\pi}_i\right]\right] = E\left[V\left[\sum_{i\in A}\frac{X_i}{\hat{\pi}_i}|\hat{\pi}_i\right]\right] + V\left[\sum_{i\in A}\frac{\pi_i}{\hat{\pi}_i}\right]$$

$$= E\left[\sum_{i\in A}\frac{\pi_i(1-\pi_i)}{\hat{\pi}_i^2}\right] + \sum_{i\in A}\pi_i^2 V\left[\frac{1}{\hat{\pi}_i}\right] + \sum_{i\in A}\sum_{j\neq i\in A}\pi_i\pi_j C\left[\frac{1}{\hat{\pi}_i},\frac{1}{\hat{\pi}_j}\right]$$

$$= \sum_{i\in A}\pi_i(1-\pi_i)E\left[\frac{1}{\hat{\pi}_i^2}\right] + \sum_{i\in A}\pi_i^2 V\left[\frac{1}{\hat{\pi}_i}\right] + \sum_{i\in A}\sum_{j\neq i\in A}\pi_i\pi_j C\left[\frac{1}{\hat{\pi}_i},\frac{1}{\hat{\pi}_j}\right]$$

$$\cong \sum_{i\in A}\frac{(1-\pi_i)}{\pi_i} + \sum_{i\in A}\pi_i V\left[\frac{1}{\hat{\pi}_i}\right] + \sum_{i\in A}\sum_{j\neq i\in A}\pi_i\pi_j C\left[\frac{1}{\hat{\pi}_i},\frac{1}{\hat{\pi}_j}\right].$$

Using this approximation, a plug-in estimate of $V\left[\hat{T}\right]$ is

$$\hat{V}\left[\hat{T}\right] = \sum_{i\in A_C}\frac{(1-\hat{\pi}_i)}{\hat{\pi}_i^2} + \sum_{i\in A_C}\hat{V}\left[\frac{1}{\hat{\pi}_i}\right] + \sum_{i\in A_C}\sum_{j\neq i\in A_C}\frac{\hat{\pi}_i\hat{\pi}_j}{\hat{\pi}_{ij}}C\left[\frac{1}{\hat{\pi}_i},\frac{1}{\hat{\pi}_j}\right],$$

which still requires $\hat{V}\left[\frac{1}{\hat{\pi}_i}\right]$ and the covariance terms. Note the change back to summing over the census records rather than the whole population, and therefore the additional $\frac{1}{\hat{\pi}_i}$ and $\frac{1}{\hat{\pi}_{ij}}$ weights.

The formula has a component of uncertainty from the imputation for the measured non-response and a second component due to the uncertainty in the imputation totals.

The imputation system estimates $\frac{1}{\hat{\pi}_i}$ through a logistic model for counted verses missed households as estimated by the survey. Therefore, we can calculate $\hat{V}\left[\frac{1}{\hat{\pi}_i}\right]$ (and the covariance terms) by:

1. applying the modelling to the first 50 bootstrap replicates,
2. doing a (quick) calibration of the $\frac{1}{\hat{\pi}_i}$ weights to the total number of individuals and households in that replicate at some higher level of aggregation where totals are published,
3. using the variation in the weights attached to the census file across the bootstrap replicates to get a simple estimate of $\hat{V}\left[\frac{1}{\hat{\pi}_i}\right]$ and $\hat{C}\left[\frac{1}{\hat{\pi}_i}, \frac{1}{\hat{\pi}_j}\right]$, or $\hat{V}\left[\sum_{i\in A_C}\frac{1}{\hat{\pi}_i}\right]$ directly.

This will give a relatively quick method to quantify the variability of a total (or cell) on a table.

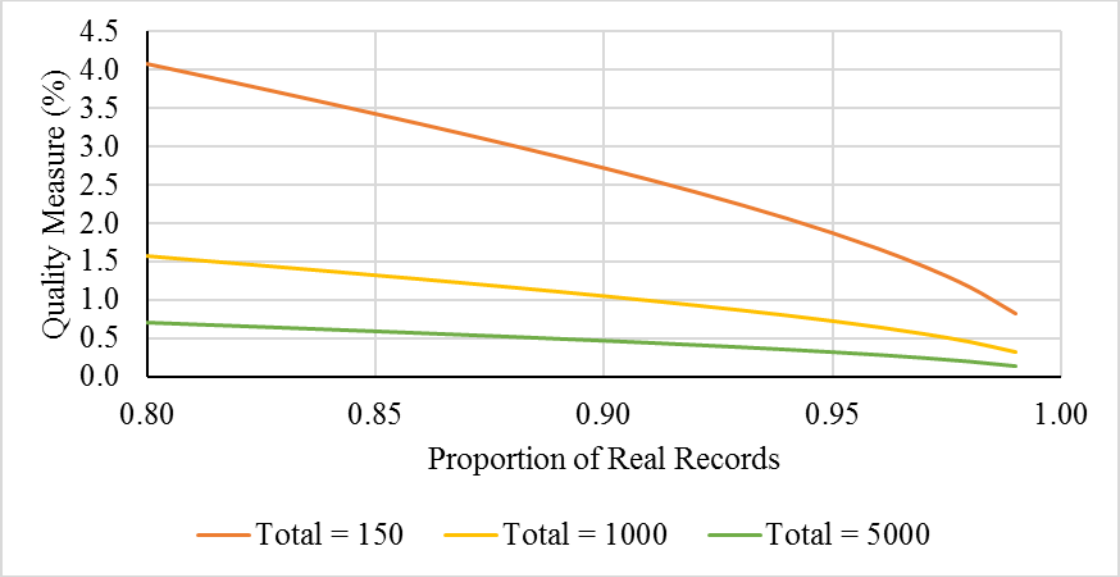
4.4) Simplified 'Quality Measure' for an Output Table

It would be desirable to remove replication from calculations relating to the imputation system. To achieve this, assume that due to the size of the coverage survey, and the aggregate level of the modelling, the terms $\hat{V}\left[\frac{1}{\hat{\pi}_i}\right]$ and $\hat{C}\left[\frac{1}{\hat{\pi}_i}, \frac{1}{\hat{\pi}_j}\right]$ will be negligible, resulting in $\hat{V}\left[\hat{T}\right] \cong \sum_{i\in A_C}\frac{(1-\hat{\pi}_i)}{\hat{\pi}_i^2}$. Then by approximating $\hat{\pi}_i$ for each record with the overall proportion of 'real' records on an output table given by $p_A = \frac{C_A}{T_A}$, we can report a quality measure for a table containing C_A census records and a published total T_A as

$$\frac{100}{T_A}\sqrt{\frac{1-p_A}{p_A^2}C_A} = \frac{100}{T_A}\sqrt{\frac{T_A - C_A}{p_A}}.$$

In other words, we just treat the count of imputed records on a table as the expectation of a Poisson random variable, and therefore its variance is equal to the count, although still weighted up from C_A to T_A. Here, the quality measure is reported relative to the size of the published count, as it should communicate uncertainty in that published total. On an Output Area table with a total size of 150 and 10% imputed, the measure would be 2.72%. This appears small but in fact the uncertainty really relates only to the 15 imputed records; so it implies the 95% confidence interval on the imputed number would be 15 ± 8. Figure 1 tracks the measure for different sized output tables and the level of imputed records.

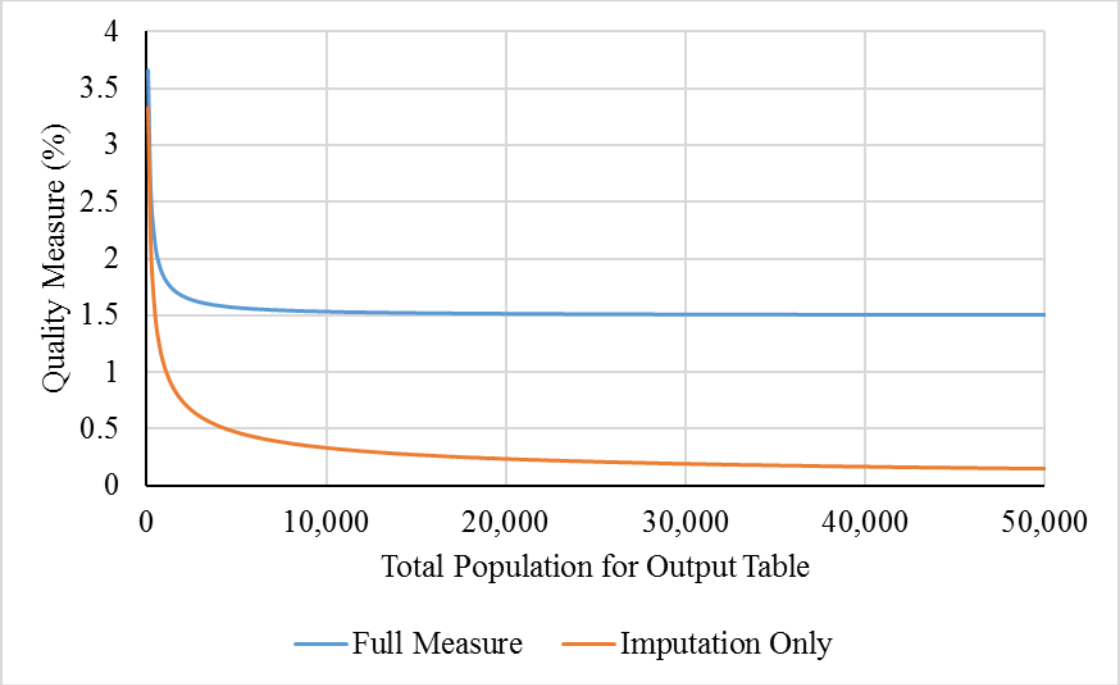
Figure 1: Quality Measure by Level of Imputation for Output Tables with Totals 150, 1000, and 5000



Reviewing the quality assurance packs from 2011, it looks as if the measure will be too low, especially as when we aggregate the quality measure for total on an output table should reflect the measured uncertainty for the published total. In other words, on higher-level tables the component due to uncertainty in the total will dominate. As a rough approximation[8], for a Local Authority with 200,000 population and around 90% response, the variance is around $3{,}000^2$. If we compute the total summing across the 18 reported age-groups, it implies the correlation would be around 0.3; we cannot ignore those correlation terms. Treating that correlation as global, and breaking down the total variance, we get that $\hat{V}\left[\frac{1}{\hat{\pi}_i}\right] \cong 0.03043^2$. Using this for an output area with 150 increases the measure from 2.72% to 3.11% (15 ± 9), while for an output area with 1,000 it increases from 1.05% (see Figure 1) to a more reasonable 1.83% (see Figure 2). Figure 2 shows that as the size of the output area increases, using these approximations results in the measure being dominated by the total uncertainty, and tending to the value 1.5%, which is the relative standard error for the total population. Figure 2 also demonstrates that on small outputs, the simple measure just based on imputation counts may be suitable, and for more aggregated areas it is sufficient to imply the area inherits the uncertainty of its parent Local Authority.

Figure 2: Quality Measure by Output Total for the 'Simple' Measure (Figure 1) and the Measure Adjusted for Total Uncertainty

[8] The approximation is based on the 2011 Census quality report for Southampton Local Authority.



Such an approach has intuitive appeal. It would improve on the information provided to users in 2011, without a large increase in computation. However, there are numerous simplifications, which need testing by at least replication on a portion of the 2011 Census database. Some results from this may already exist due to the post-2011 work by Sexton and Brown.

5) Concluding Remarks

Given the use of bootstrapping for 2011 Census, there is no justification for not doing so in 2021. Bootstrapping offers flexibility to build an estimation system with multiple components operating on different, but typically over-lapping, components of the national sample. It is not a panacea and careful implementation is required with the detailed stratification of the current coverage survey design. The extension to empirical confidence intervals is desirable, but implementing the reliable BCa approach requires considerable replication in terms of the bootstrap and the addition of the jackknife. This will likely be only possible for a (pre-) specified set of estimates, such as age-sex by Local Authority.

The proposal in Section 4 relating to the imputed database will need testing before it can be considered a useful summary measure to accompany output tables. Just accounting for imputation looks too low, and there is the need to include some uncertainty due to the estimated totals. Getting a simple measure for that, without applying full replication, requires further simplifying assumptions.

References

Baillie, M., Brown, J., Taylor, A., and Abbott O. (2011) Variance estimation. Titchfield: ONS [cited 27/03/20]. Available from

https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howweplannedthe2011census/independentassessments/independentreviewofcoverageassessmentadjustmentandqualityassurance/imagesvarianceestimationv1tcm774049_tcm77-224787(1).pdf.

Brown, J. J. (2000) Design of a census coverage survey and its use in the estimation and adjustment of census underenumeration. *PhD Thesis*. University of Southampton, Southampton.

Brown, J., Abbott, O., and Smith, P. A. (2011) Design of the 2001 and 2011 census coverage surveys for England and Wales. *Journal of the Royal Statistical Society Series A*, **174** pp881-906.

Brown, J. J., Sexton, C., Abbott, O., and Smith, P. A. (2019) The framework for estimating coverage in the 2011 Census of England and Wales: Combining dual-system estimation with ratio estimation. *Statistical Journal of the IAOS*, **35**, pp481–499.

Efron B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, pp1–26.

Efron, B. & Tibshirani. R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall: New York.

Gross, S. (1980) Median estimation in sample surveys. *Proceedings of Section for Survey Research Methods*, American Statistical Association, pp181–184.

Kabzinska, E., Smith, P, and Berger Y. (2017) Empirical likelihood approach to census coverage estimation. *New Techniques and Technologies for Statistics 2017* [cited 27/03/2020]

https://www.conference-service.com/NTTS2017/documents/agenda/data/full_papers/full_paper_152.pdf

Rubin, D. (1987) *Multiple imputation for nonresponse in surveys*. Chichester: Wiley.

Steele, F., Brown, J., and Chambers, R. (2002) A controlled donor imputation system for a one-number census. *Journal of the Royal Statistical Society Series A*, **165**, pp495–522.

Wolter, K. M (2007). *Introduction to Variance Estimation*. 2nd Edition. New York: Springer.