Performance of 2021 Coverage Adjustment Method with Lower Response and Contingencies

MARP (October 2020)

Rajni Sandhu and Kirsten Piller

1. Introduction

The census coverage adjustment amends the unit level census database to make it consistent with the population estimates derived from the coverage estimation process, so robust estimates can be obtained for lower level geographies. Historically the adjustment has been made to account for census under coverage of both households and people. This paper follows two strategy papers on adjustment that were presented to the panel in 2018 and 2019. It covers the use of the Combinatorial Optimisation (CO) method for adjustment with lower levels of person and household response, and contingencies that may be required if the level of response impacts the performance of CO. The 2021 adjustment system will be made up of two stages:

Stage 1.
Impute missed households (and persons within them) using the following steps.

i) Derive integer benchmarks for population and household totals by key demographic characteristics that represent the missed households and persons, using the coverage weights provided by the coverage estimation system.
ii) Select donor households using the Combinatorial Optimisation method, ensuring the benchmarks in i) above are maintained as closely as possible.
iii) Place the donor households in an appropriate postcode.

Stage 2.
Impute characteristic variables for the persons and households imputed in Stage 1 using CANCEIS (CANadian Census Edit and Imputation System).

The first strategy paper (Whitworth et al, 2018) described some initial work that demonstrated proof of concept for the new methods in 5 estimation areas (EA, groupings of one or more local authority). It referred to Oguz and Abbott (2016), which demonstrated that overall the benchmarks for age group by sex (at local authority level) and other key characteristics (at estimation area level) were better met using the CO method than when using the 2011 method. It also described some subsequent analysis showing that the benchmarks were also better met in a two-stage approach (stated above) compared to a three-stage approach (whereby missed people are first imputed into counted households).

The second paper (Whitworth et al, 2019) provided more detailed evidence on the implementation of the strategy, including the performance of the CO method in a two-stage approach and investigation of properties of the CO outputs. It covered the performance of the CO method when using benchmarks for more detailed categorisation of variables at estimation area level. For most benchmark variables a good fit was achieved with more detailed categories. It also presented an initial analysis of CO with lower response, which has been continued in this paper. Plans were also outlined for using administrative data to provide additional evidence to inform the methods.

The purpose of the analysis in this paper was to assess the performance of CO when the person and household response rates were lower than expected. Simulations were set up with scenarios where different amounts of persons and households were removed from the census database for various local authorities. CO assessment measures were used to compare the performance of CO from each

scenario to the performance with the original 2011 data. Section 2 of this paper presents the analysis; it gives an overview of the method and measures used to assess the performance and the results. A key change for 2021 will be that the adjustment will make use of local authority (LA) census population estimates for a wider range of key demographic characteristics and breakdowns produced by the estimation system. Section 3 describes and discusses the contingencies that will be considered when the performance for a local authority could be improved, and Section 4 covers the next steps for investigating the performance of CO.

2. Analysis

2.1. Method

For this research, the CO method was applied with 2011 Census data as the donor pool and constrained to local authority level benchmarks. These benchmarks were produced using post-adjustment 2011 Census data, since 2011 coverage estimates were only available at EA level and could not be broken down to local authority level.

The number of missing persons and households was increased in order to understand how CO performs and whether a contingency method will be necessary if the response rate is lower than expected. A contingency method to consider is the imputation of persons into households before implementing CO for wholly missed households.

Two scenarios were considered with increasing levels of missing persons and households to compare back to the 2011 results. These scenarios were:

Scenario	Count missing	Count removed
1	1.5 x count missing in 2011	0.5 x count missing in 2011
2	2 x count missing in 2011	1 x count missing in 2011

In this analysis, an assumption is made that the proportion of persons missing is uniform with respect to the proportion of households missing in each scenario. The count missing was calculated by comparing the total number of households/persons in the 2011 Census data (before adjustment) with the total number in the post-adjustment data.

For each scenario the required number of households were removed using simple random sampling, but within a limit. The samples for each simulation were taken from a list of households with a low probability of being counted in the census, by only considering households with probability in the lower quartile. By removing households, persons have also been removed, so the number of persons removed with those households are subtracted from the total number of persons to be removed. The remaining number of persons are removed by only considering persons with high probability of being missed in the census. A simple random sample was taken from a list of persons with probability in the upper quartile. Only persons in households of size 2+ were considered for removal and the method ensured at least one person from each household remained in the donor pool to avoid further removing households. Although the number of persons removed in wholly missed and counted households was not controlled for separately, after removing whole households there was still a large proportion of persons remaining to be removed. This proportion was also consistent across the simulations within each scenario and LA.

The analysis was carried out for three local authorities with different characteristics. Waltham Forest (00BH) is a local authority in outer London with a large population in proportion to the number of households and a low response rate, High Peak (17UH) in the East Midlands has a higher response rate and a smaller population, and West Somerset (40UF) has the highest response rate and lowest

population. The number and percentage of missing persons and households for each local authority are contained in tables 1, 3 and 5 in the report.

20 simulations were run for each scenario. As a different sample of persons and households was removed in each simulation, the shortfall (difference between the number of persons/households counted and the benchmark) is slightly different for each simulation. The variation in shortfall between simulations is larger in scenario 2 than in scenario 1 since a larger sample of persons and households are being removed.

The demographic characteristics used for the benchmarks are:

- Age-sex groups: 35 categories (5-year age groups)
- Ethnicity: 5 categories
- Activity last week: 5 categories
- Household size: 1, 2, 3, 4, 5+
- Hard to count index: 5 possible categories (index 1 and 2 for High Peak and West Somerset, index 3 only for Waltham Forest)
- Tenure: 5 categories

The CO method completed 50 runs for each simulation so there are (50*20=) 1000 runs for each scenario. The 50 CO runs vary by their starting point, which affects the final selection of households obtained for each run.

The Overall Total Absolute Error (OTAE) is used to assess how well the CO method has met all benchmark variables in each run. It is a sum of the Total Absolute Errors (TAE) for each of the benchmark variables. The TAE for each variable is the sum of the absolute differences across categories between observed counts from the final selection of households by the CO method and the expected counts (the benchmark to be met).

$$\text{TAE} = \sum_{ij} |O_{ij} - E_{ij}|$$

Where:

O_{ij} is the observed count for category j in variable i (observed value)
E_{ij} is the expected count for category j in variable i (expected value)

Two additional measures presented in this report are used to assess the performance of the CO method. The maximum duplicates measure is the maximum number of times a household has been duplicated in the final selection of households produced by the CO method. Proportion of unique households is calculated as:

$$Proportion\ unique = \frac{Number\ of\ unique\ households\ in\ final\ selection}{Total\ number\ of\ households\ imputed}$$

The maximum duplicates and proportion of unique households provide measures of the variation of households in the final selection produced by CO.

Since this investigation involves removing persons from counted households without changing the benchmarks, there are more small households and fewer large households in the donor pool than before. This has resulted in an overcount of one of the household size categories in High Peak and Waltham Forest. This causes negative shortfall for some of the simulations, where the census count for a variable category is higher than the estimate, therefore the number of households that need to be imputed is negative. Households are not removed as part of the adjustment process however, CO

balances the excess households from one category by undercounting households in the other categories. This gives a larger OTAE as the minimum OTAE will be twice the value of the negative shortfall. A small negative shortfall will not have much influence on the results, but a larger negative shortfall might. Even though negative shortfall has occurred in this analysis, it is unlikely that it will occur in 2021 as the coverage estimation method that will be used ensures that the estimates will not be less than the census count for the household size variable. Any error that has been caused by negative shortfall has been removed from the OTAE in this report.

2.2. Results

00BH (Waltham Forest)

Table 1 – response rate and number of missing persons/households in each scenario for 00BH

		2011:	Scenario 1:	Scenario 2:
Households	Percentage counted	92.78%	89.18%	85.57%
	Count missing	6,994	10,491	13,988
	Count to remove		**3,497**	**6,994**
Persons	Percentage counted	89.96%	84.94%	79.92%
	Count missing	25,906	38,859	51,812
	Count to remove		**12,953**	**25,906**

The large OTAE seen in Figure 1 below for this LA is not a true reflection of the error expected for this LA. It is likely due to the benchmark being calculated from the adjusted 2011 Census data, as this LA did not control for household size as part of the 2011 adjustment. Also, this LA is larger, and the shortfall is larger (Table 1), so the error is relative to this. This LA is one of a few that entirely makes up one EA, so 2011 coverage estimates were available and the OTAE with these as benchmarks is much lower (around 1,500). Most of this error is contained in the household size variable (see Appendix 1 for more detail) and there is minimal error in the other benchmark variables. This is a common result from the CO method as can be seen in the results presented in previous papers. The OTAE has increased by about 500 in scenario 1 and a further 500 in scenario 2. For all simulations, the average error is higher than for the 2011 scenario, but the range is similar. Negative shortfall increased the OTAE in this LA, but it has been removed in the results and was small compared to the final values for OTAE so did not have a huge influence on the results.

Figure 1 - boxplot to compare average OTAE for each simulation by scenario

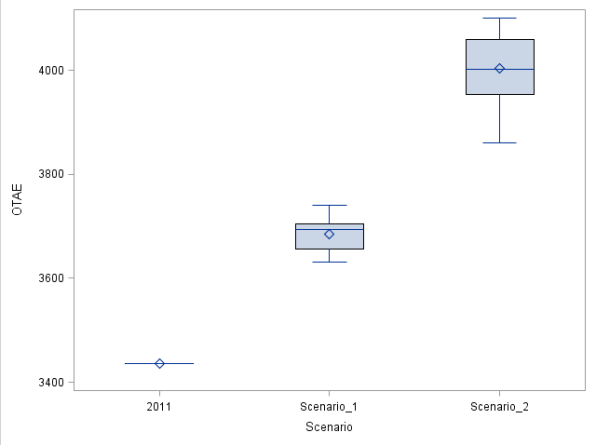


Table 2 – summary of OTAE, maximum duplicates and proportion of unique households from all runs for each scenario

Measure	Scenario	Min	Max	Range	Mean
OTAE	2011	3122	3690	568	3436.82
	1	3311	4095	784	3685.063
	2	3608	4420	812	4003.275
Maximum duplicates	2011	52	109	57	82.72
	1	47	134	87	72.407
	2	48	159	111	77.501
Proportion unique households	2011	0.0902	0.0999	0.0097	0.0955
	1	0.1208	0.1344	0.0136	0.1282
	2	0.1436	0.1567	0.0132	0.1496

Table 2 summarises the results from all runs in all simulations for each scenario, whereas Figure 1 displays points representing the average across all 50 CO runs in each simulation (one point for the 2011 scenario). See Appendix 2 for results from all simulations displayed separately. Table 2 shows that as the error increases with more missing persons and households, the maximum number of duplicates decreases in scenario 1 and increases slightly in scenario 2, but the proportion of unique households in the selection increases in both scenarios. The maximum duplicates measure gives an indication of the quality of the selection of households in each simulation as a high number of duplicates is undesirable. Even though on average households are duplicated more often in scenario 2, there appears to be more variation in the final selection of households as the proportion of unique households has increased.

There is an additional diagnostic that is calculated as the OTAE/number of households imputed (hh). Even though the error has increased with more missing persons and households, the OTAE/hh has decreased for each scenario from an average of 0.49 in the 2011 scenario, down to 0.35 in scenario 1 and 0.29 in scenario 2. This shows that even though the average error has increased by approximately 600 from the 2011 scenario to scenario 2, it is small compared to the increasing number of households being imputed. Generally, a lower value is considered better because it means that the error is low compared to the number of households being imputed.

17UH (High Peak)

Table 3 – response rate and number of missing persons/households in each scenario for 17UH

		2011:	**Scenario 1:**	**Scenario 2:**
Households	Percentage counted	96.68%	95.02%	93.36%
	Count missing	1,294	1,941	2,588
	Count to remove		**647**	**1,294**
Persons	Percentage counted	96.11%	94.17%	92.22%
	Count missing	3,496	5,244	6,992
	Count to remove		**1,748**	**3,496**

Scenario 2 for this LA had negative shortfall that contributed more than 100 to the OTAE in most simulations which was a large proportion of the average OTAE for this scenario. Any error caused by negative shortfall was removed from the OTAE in the results, but it is not possible to determine how this has affected the maximum duplicates and proportion of unique households. Even after the error caused by negative shortfall has been removed, most of the OTAE shown is present in the household size variable.

All simulations in each scenario have a lower average OTAE. This may be because, since more households are required to meet the benchmarks, there are more possible selections that can be used to meet these results. The analysis was repeated for another local authority with similar characteristics and from the same EA and similar results were observed.

Figure 2 - boxplot to compare average OTAE for each simulation by scenario

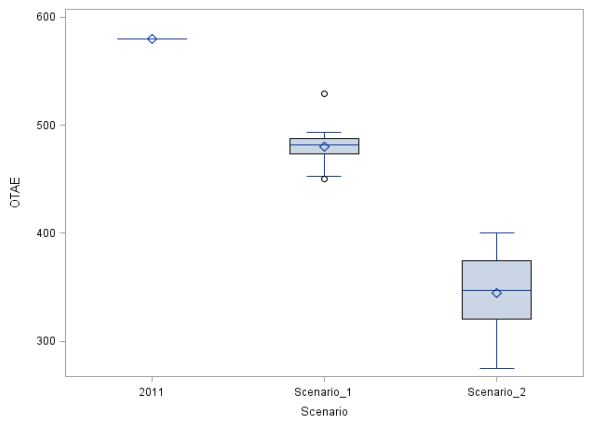


Table 4 – summary of OTAE, maximum duplicates and proportion of unique households from all runs for each scenario

Measure	Scenario	Min	Max	Range	Mean
OTAE	2011	559	602	43	580.08
	1	307	702	395	480.206
	2	137	587	450	344.514
Maximum duplicates	2011	4	78	74	10.12
	1	17	195	178	105.527
	2	66	219	153	111.694
Proportion unique households	2011	0.0317	0.9312	0.8995	0.8017
	1	0.0268	0.6579	0.6311	0.0377
	2	0.0301	0.0483	0.0182	0.0391

Table 4 summarises the results from all runs in all simulations that have been run for each scenario whereas Figure 2 displays points representing the average across all 50 CO runs in each simulation (see Appendix 2 for results from all simulations displayed separately). The average OTAE after removing persons and households is much lower than the 2011 scenario, and the range is also wider. When comparing scenario 1 with scenario 2, the average OTAE is similar but the range of results is wider for scenario 2. This is likely related to the differences between the samples of households and persons removed and the larger shortfall between the responses and benchmarks in each simulation rather than the just performance of CO.

The maximum duplicates are generally higher in scenario 1 and 2 compared with the 2011 scenario, which corresponds to a lower proportion of unique households. There appears to be a link between OTAE and the number of duplicates in the selection of households as, when there is a lower OTAE, the number of duplicates increases. For the 2011 scenario, the CO method produced selections where on average the proportion of unique households in the selection was 0.8 however, this dropped to less than 0.4 in scenario 1 and 2.

Even though the OTAE in scenario 2 is only slightly lower than scenario 1, the average OTAE/hh decreases from 0.25 in scenario 1 down to 0.13 in scenario 2. This is down from 0.45 in the 2011 scenario.

40UF (West Somerset)

Table 5 – response rate and number of missing persons/households in each scenario for 40UF

		2011	**Scenario 1:**	**Scenario 2:**	**Scenario 3:**
Households	Percentage counted	97.62%	96.43%	95.24%	88.10%
	Count missing	372	558	744	1,860
	Count to remove		**186**	**372**	**1,488**
Persons	Percentage counted	97.15%	95.73%	94.30%	85.75%
	Count missing	949	1424	1898	3,796
	Count to remove		**475**	**949**	**2,847**

Scenarios 1 and 2 were run for West Somerset, but since the OTAE was low for the 2011 scenario, there was not a clear change to the results for these scenarios. Therefore, a further scenario was considered for this area. Scenario 3 has five times as many missing persons and households as the 2011 scenario.

Figure 3 - boxplot to compare average OTAE for each simulation by scenario

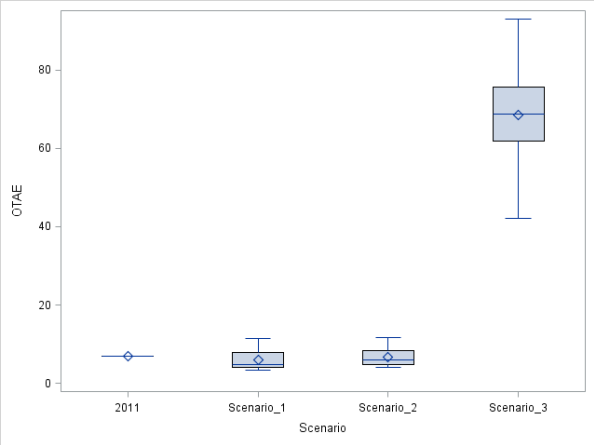


Table 6 – summary of OTAE, maximum duplicates and proportion of unique households from all runs for each scenario

Measure	Scenario	Min	Max	Range	Mean
OTAE	2011	0	16	16	6.9
	1	0	25	25	6.045
	2	0	22	22	6.587
	3	14	218	204	68.519
Maximum duplicates	2011	2	46	44	23.64
	1	2	57	55	14.098
	2	3	65	62	20.049
	3	26	238	212	96.531
Proportion unique households	2011	0.5484	0.9731	0.4247	0.7126
	1	0.3835	0.9659	0.5824	0.7991
	2	0.4234	0.9422	0.5188	0.7239
	3	0.0263	0.4747	0.4484	0.1261

Figure 3 shows that the average OTAE has increased in scenario 3 but from Table 6 the average maximum duplicates also increased. This is different to what has been assessed in High Peak and Waltham Forest where an increase in OTAE is linked to a decrease in maximum duplicates. This could be because the OTAE and maximum duplicates were close to 0 in the other scenarios, so it would not be possible for either of the OTAE or the maximum duplicates to decrease while the other increases. Also, in scenario 3 the distribution of household and person characteristics available for the CO method is reduced, so it is likely to be more difficult to find a selection that perfectly meets the constraints without duplicating many households. As well as the maximum duplicates increasing in scenario 3 the proportion of unique households decreases from 0.71 in the 2011 scenario down to 0.13 in scenario 3.

The OTAE/hh shows a constant decrease as more persons and households are removed. From 0.019 in the 2011 scenario, it decreases to 0.011 in scenario 1 and 0.009 in scenario 2. Therefore, the error in proportion to the number of households being imputed decreases as more persons and households are removed. However, it increases in scenario 3 to an average of 0.037. While this is higher than the average for the 2011 scenario, it is still a very small value compared to the results for other local authorities.

2.3. Discussion

As discussed in previous papers, the CO method performs well for imputing persons and households in one stage based on response rates in 2011 when compared to the method used in 2011. This analysis has provided some information on how the CO method will perform at local authority level if the response rates are lower than they were in 2011.

There are a range of results for all three of the areas assessed. Waltham Forest performed as expected by showing increased OTAE as response rate decreased, but this was small in comparison to the increased number of persons and households missing. CO appeared to duplicate fewer donors in scenario 1 and 2 but the maximum duplicates did increase slightly in scenario 2 as the donor pool became more limited. For High Peak the average OTAE decreased as response rate decreased, but the maximum duplicates increased as CO struggled to find enough donors to match the benchmarks from the smaller donor pool. Finally, West Somerset did not have a clear change in the results for up to double the number of missing persons and households. However, for five times the number of missing responses in 2011 there was an increase in OTAE and maximum duplicates, again as the donor pool became more limited and the shortfall was larger.

Based on the three areas assessed, while there is no clear pattern in the results for a decrease in response rate, CO still appears to perform well. The 2021 adjustment approach will only use CO to impute wholly missing households (to cover both wholly missing households and persons missed from counted households). There may be cases where we will have to consider imputing persons into counted households before running the CO method on only wholly missing households. This is discussed more in the following section.

3. Contingencies

There are some patterns between the levels of error, duplication and response for an area in this and other research, but there have always been some exceptions. How CO performs will come down to the detailed make up of each area's household and person responses. Based on the analysis above, there may be some unexpected results for different areas, but contingencies will be put in place to deal with these.

Aside from the measures that have been used to assess the performance of CO in this paper, other measures that will be considered are:

- The distribution of duplicates for the final selection of households.
- The breakdown of error across variables and categories (see Appendix 1 where household size has been presented in this way):
    - for the final selected run and
    - as an average of the 50+ runs of CO, including the variation across the runs in relation to the size of the shortfall.
- Whether the adjusted census totals are contained within the confidence intervals of the coverage estimates.
- Chi-Squared test for the difference between the expected (benchmark) and observed (adjusted) counts in the final selection.

3.1. Initial contingencies

The following contingencies will be considered first if the performance of CO is concerning for an area:

- The adjustment process allows CO to produce 50 runs or more for each area, i.e. there will be 50 selections of households available for each area. This is done to explore the solution space to find a global minimum. The selection of households in each run will vary in their characteristics so if the selected run is not appropriate one of the other runs may work better. This may be even more important with lower response as there is likely to be more variation in the final selections when there is a larger shortfall for the benchmarks. The method of determining the 'best' run has not been described in this paper but will essentially choose the run that minimises

the OTAE, the TAE in the priority benchmark variables (age-sex group and tenure) and the level of household duplication.

- It is possible to change how CO selects households by adapting the simulated annealing parameters, which will have default set of values for each area to begin with. We will be carrying out more investigations on how this impacts the selection of households by CO. The parameters affect aspects of the CO method such as how many times to swap households in the selection (iterations) before accepting the OTAE, the probability a record is to be accepted into the selection and how quickly that probability changes (it reduces during iterations).
- If a high number of duplicates is the main concern for an area, a limit could be included for the maximum number of duplicates for a household. This may increase the overall error, but if the OTAE is low to begin with then it may not be a concerning increase. Incorporating a limit would also considerably increase processing time.
- We may also consider optimising the selection of households by including donors from outside of the donor pool, such as a neighbouring LA. For example, if CO is struggling to find large households to impute (and is duplicating too many households) with lower response, this contingency would provide more variety of households of each household size in the donor pool. This option has not been explored yet to determine if it is feasible or efficient, and it would require a method to determine how to select an area with households that would provide the desired characteristics for improving the final selection.

3.2. Imputation of persons into counted households

If the CO method is unable to find a selection of households that meets the constraints well enough, it may be necessary to incorporate some of the 2011 coverage adjustment methodology where persons were first imputed into counted households (see Appendix 3 for more detail). Unless there are some direct indicators that there is a lower level of responses from persons within households and this is affecting the performance of CO for selecting households, the person imputation will be considered after the contingencies given in 3.1. These contingencies would be simpler to implement and would not significantly increase the process run time.

A contingency strategy for also imputing persons would involve:

Stage 1. Impute missed persons from counted households using 2011 methods.
Stage 2. Impute missed households using CO.
Stage 3. Impute characteristic variables for the persons and households imputed in Stages 1 and 2.

The first strategy paper (Whitworth et al, 2018) presented the improvements in the performance of CO for the whole adjustment process, although the use of CO post-person imputation still showed improved results over the 2011 method. More information on the use of CO in a three-stage approach (at EA level) is available in Oguz and Abbott (2016).

However, the results in this paper haven't necessarily suggested that person imputation would be required. This contingency will be considered by area, as we will come across areas that have different distribution of characteristics and levels of response to the areas that we have been able to test. Therefore, the performance of CO will vary between different areas. Determining when person imputation might be required will depend on a few factors, and it may not necessarily improve how well the adjusted census meets the benchmarks. The TAE in household size tends to be the largest of all the benchmark variables in adjustment, no matter the approach (Appendix 1). If the TAE is particularly large, and mostly concentrated in the larger household size category this may indicate

more of a need to impute persons first for an area. However, this may also be improved by widening the donor pool to include a neighbouring LA instead.

If this three-stage approach were implemented, an additional process of estimating the number of persons in missed households separately from the number of persons in counted households is required as coverage estimation will only provide population estimates for the total number of people in the population. The 2011 adjustment method used model estimated probabilities (inverted to be coverage weights) to separate out the estimates. The models for each estimation area provided an estimated probability of response for every person and household response. The fit of these models varied by area and required changes during processing in 2011 and may not fit as well at LA level for 2021.

Also, when persons are imputed into counted households, they require their relationship to other household members imputed. This will add processing time to the post-adjustment item imputation process and as it is a difficult variable to impute, may result in low quality imputation.

If an area requires person imputation in 2021, the 2011 methodology can be used as it was to run models for each area as part of adjustment. However, there may be outputs from the coverage estimation methods for 2021 that could feed into this instead and would cut down on contingency processing time for adjustment. The estimation models will be run at a more aggregate level than in 2011 so they will likely provide better information than the area level models used in adjustment in 2011.

4. Next steps

1. The CO method will be run on multiple simulated unadjusted census databases from a true census population. This simulation set up will allow us to better assess the variability and bias in the different adjusted census databases.

   This work will use outputs from coverage estimation research (population estimates from simulations at local authority level), so we will be able to better assess the performance of the adjustment at local authority level. CO research so far has either been carried out at EA level with 2011 coverage estimates or using the 2011 adjusted census totals as LA level benchmarks. Also, as the outputs are from coverage estimation there will be no negative shortfall in the benchmark variables.

   In the longer term, this could feed into creating measures of uncertainty for census statistical processing. This simulation set up could also include the coverage scenarios and contingencies explored in this paper, and even look at scenarios with a much lower response for persons within counted households. It will be easier to account for different levels of response with this set up as persons and households will be selected to be included in simulation unadjusted census databases rather than being removed.

2. We will also look at incorporating contingency routing into the 2021 adjustment system and creating a strategy for when each contingency should be considered.

5. References

Oguz, S and Abbott, O., (2016), "2021 Census Coverage Adjustment Methodology", 31st Meeting of the GSS Methodology Advisory Committee, ONS.
https://gss.civilservice.gov.uk/wp-content/uploads/2013/03/GSS-MAC-31-Booklet_3.pdf

Whitworth, A, Sexton, C and North, R., (2018), "The 2021 Census Coverage Adjustment Strategy", Unpublished internal working paper for the External Review (September 2018).
https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP106-The-2021-Census-Coverage-Adjustment-Strategy.docx

Whitworth, A, Piller, K, Sandhu, R and Penn, A., (2019), "The 2021 Census Coverage Adjustment Strategy Update", Unpublished internal working paper for the External Review (October 2019).
https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP122-The-2021-Census-Coverage-Adjustment.docx

Appendix 1: OTAE in household size variable

Testing the CO method has shown that the OTAE doesn't always reach 0 and when there is a large error, most of it is present in the household size variable. This is a feature of running the adjustment process with household size as a 5-level variable. While this is a seemingly undesirable result, by including household size as a 5-level constraint, the CO method can control for it as well as possible and therefore performs better than when household size is collapsed to fewer levels. The example below investigates the results for household size in EA OL13WALT (Waltham Forest). Table 1 shows the results from the method used in 2011. Table 2 shows the result from Oguz and Abbott (2016) where household size had a TAE of 14 for the household size variable when it was run at 2 levels. This shows that when looking at this result in more detail, the error in the household size 2+ level was hiding the error that is present in the households of larger sizes as they were not controlled for. Similar results have been seen for EAs IL03GREE (Greenwich, equivalent to 1 LA) and EM05NSDE (North and South Derbyshire, equivalent to 7 LAs).

Table 1 – Error in the household size variable from the three-stage approach used in 2011 with household size as a 2-level variable.

Household size (2 categories)	Adjusted census total	Coverage estimate	Difference
1	28576	28837	261
2+	68343	68082	-261
		TAE:	**522**
Split household size into 5 categories:			
1	28576	28837	261
2	24279	25890	1611
3	16833	16452	-381
4	14492	13741	-751
5+	12739	11999	-740
		TAE:	**3744**

Table 2 – Error in the household size variable when CO is run after persons have been imputed into households using the 2011 method with household size as a 2-level variable.

Household size (2 categories)	Adjusted census total (Average of 100 CO runs)	Coverage estimate	Difference
1	28830	28837	-7
2+	68089	68082	7
		TAE:	**14**
Split household size into 5 categories:			
1	28830	28837	-7
2	24331	25890	-1559
3	16370	16452	-82
4	14492	13741	751
5+	12896	11999	897
		TAE:	**3296**

Table 3 – Error in the household size variable when CO is run as part of the two-stage approach with household size as a 2-level variable.

Household size (2 categories)	Adjusted census total (Average of 100 CO runs)	Coverage estimate	Difference
1	28837	28837	0
2+	68082	68082	0
		TAE:	**0**
Split household size into 5 categories:			
1	28837	28837	0
2	24871	25890	-1019
3	16364	16452	-88
4	13833	13741	92
5+	13014	11999	1015
		TAE:	**2214**

Table 4 - Error in the household size variable when CO is run as part of the two-stage approach with household size as a 5-level variable.

Household size (5 categories)	Adjusted census total (Average of 100 CO runs)	Coverage estimate	Difference
1	28101	28837	-736
2	25890	25890	0
3	16452	16452	0
4	13741	13741	0
5+	12735	11999	736
		TAE:	**1472**

Appendix 2: OTAE for each simulation

Figures 1 and 2 show the OTAE for the 50 runs in each simulation and the 50 runs in the 2011 scenario for High Peak. Figures 3 and 4 show the same results for Waltham Forest. Figure 5 shows the results for scenario 3 and the 2011 scenario for West Somerset. The results for scenario 1 and 2 for West Somerset are not included as the OTAE for many of the runs were 0 or very small for each of the simulations.

Figure 1 - boxplot of OTAE from the 2011 scenario and 20 simulations from scenario 1 for 00BH

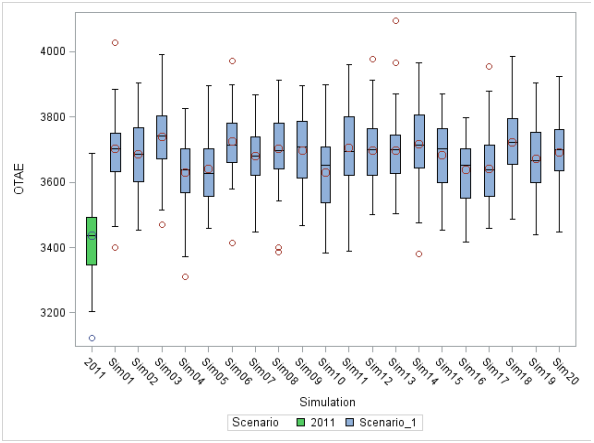


Figure 2 - boxplot of OTAE from the 2011 scenario and 20 simulations from scenario 2 for 00BH

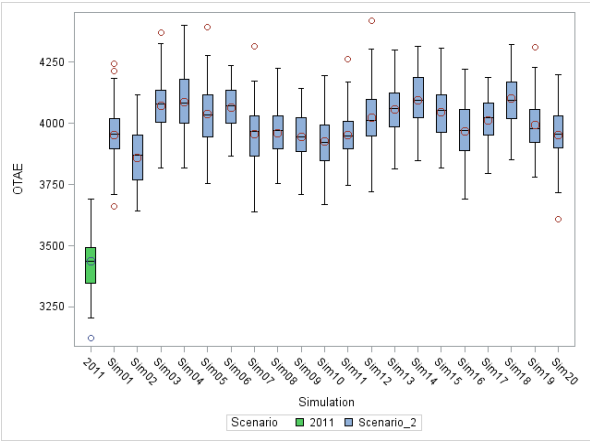


Figure 3 - boxplot of OTAE from the 2011 scenario and 20 simulations from scenario 1 for 17UH

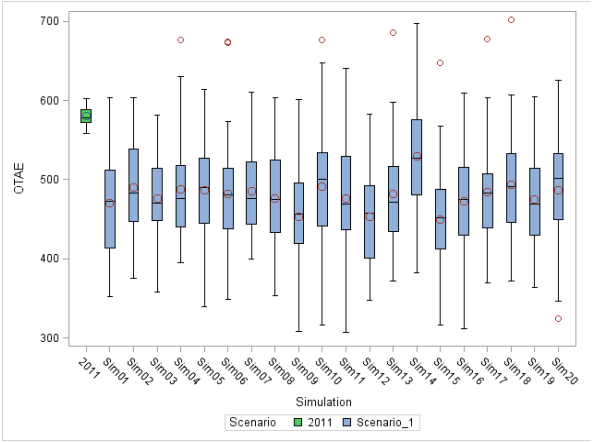


Figure 4 - boxplot of OTAE from the 2011 scenario and 20 simulations from scenario 2 for 17UH

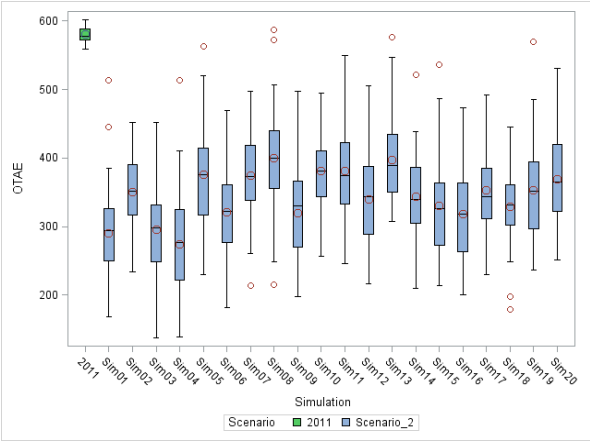
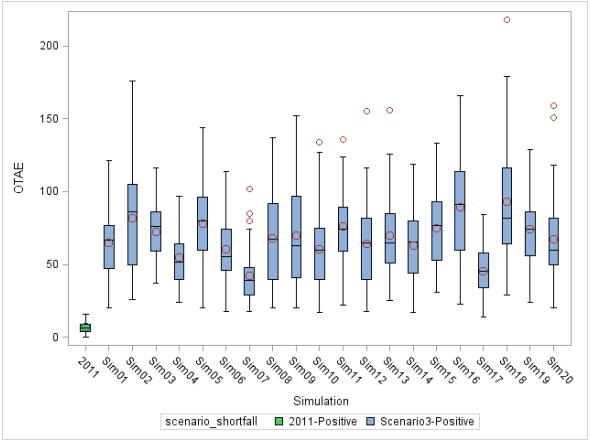


Figure 5 - boxplot of OTAE from the 2011 scenario and 20 simulations from scenario 3 for 40UF



Appendix 3: 2011 method to impute persons

The 2011 coverage adjustment method had three stages completed separately for each estimation area (EA):

- Stage 1 - impute missed persons into counted households in the census database.
- Stage 2 - impute missed households and persons within them into the census database.
- Stage 3 - impute characteristic variables for the persons and households imputed in stages 1 and 2 using CANCEIS.

The adjustment strategy for 2021 only has stages 2 (using CO) and 3. Stage 1 could be included prior to this if imputing persons is required as a contingency for some areas in 2021. Below is a description of the method used for stage 1 in 2011.

1. Multinomial generalised logit models are fit to matched Census to CCS for each EA. These models are used to predict probabilities of response for all census person records in the EA. Two person-level models are fitted, one for children (age < 16) and the other for adults. The response variable has the following levels:

   P_1: person counted in both a census and CCS counted household

   P_2: person counted in CCS but missed from a household that was counted in the census

   P_3: person that was counted in the CCS but belongs to a household that was missed by the census

   Fixed variables are used in default child and adult models, so there is no model selection for each area. However, the system allows for a modified model to be specified if the model fits poorly.

   Default variables in the person-level models:

Adult model	**Child model**
Age-sex group	Age-sex group
Hard to count index	Hard to count index
Tenure	Tenure
Adult collapsing of household structure	Child version of household structure
Local authority	Local authority
Activity last week	Ethnicity
Marital status	Born UK
Ethnicity	Address year ago
Address year ago	
Born UK	
Intention to stay	

2. Coverage weights are calculated as the reciprocal of P_1 (the probability of a person being counted in the census), and these weights are calibrated to the coverage estimates. The calibration procedure applies raking ratio to the weights. For this, the weights are iteratively scaled to the estimates of each coverage estimate variable until convergence is achieved (minimal overall difference to coverage estimates). The variables for calibration are ordered:

activity last week, ethnicity, tenure, hard to count index and age-sex by LA. Age-sex by LA is prioritised by being the last variable to calibrate to. The calibration process will stop if convergence is not achieved within the maximum number of iterations set.

The calibrated coverage weights need to be adjusted so that they represent only persons missed from counted households. These weights are derived using the first set of calibrated weights and the remaining two components of the two fitted models.

3. The person-level data with derived weights are sorted by age-sex group, weight, OA code and postcode. The selection of persons to impute is carried out by each LA within the EA. For each age-sex group within each LA, cumulative sums of the weights and person count are calculated. Whenever the rounded difference between these sums is at least 1, the record of person at which this happens is marked to be copied however many times it takes to get the person count up to the current rounded cumulative sum of weights.

Record	Coverage weight	Cumulative sum of persons (A)	Cumulative sum of coverage weights (B)	Rounded diff (B – A)	Imputed person
1	1.2	1	1.2	0	
2	1.1	2	2.3	0	
3	2.6	3	4.9	**2 (impute this person 2 times)**	
3	n/a	4	n/a	n/a	**Y**
3	n/a	5	n/a	n/a	**Y**
4	1.0	6	5.9	0	
5	1.5	7	7.4	0	
…					

4. Persons to be imputed are placed into suitable households. For each donor, a sequence of search procedures is carried out until a suitable household is found for that donor to be placed into, with the constraints on what constitutes a suitable household for placement being relaxed with each subsequent search.