**Paper 9. The 2021 Census Coverage Adjustment**

**Context and aims**

This paper reviews development of the census coverage adjustment methodology for the 2021 Census. The adjustment amends the unit level census database so that it accounts for net under-coverage of both households and people, it uses an imputation approach with area level population estimates as benchmarks. For 2021, the adjustment strategy has been designed to address the practical difficulties that were experienced during implementation of the 2011 methods, and to make best use of the new strategy and outputs from the coverage estimation. Some key changes have been made to the methods to improve the quality and transparency of the outputs, and to increase the efficiency of the adjustment processing system.

This update paper follows the initial strategy paper presented in 2018. It summarises work undertaken to test the statistical properties of the adjustment outputs based on the new methods and outlines how administrative data can be used to enhance the evidence base for these methods. The paper refers to earlier work, providing detail of the underlying methods and describing initial (proof of concept) research. The supporting papers are provided for ease of reference but do not necessarily need to be read in full.

The panel are asked to comment on the strategy, methods and work presented, and to provide advice on the direction for development of the coverage adjustment methodology.

**Specific issues for discussion**

**Q1 Should we continue to pursue the two-stage approach?**

**Q2 Does the panel have suggestions for how to deal with a scenario where there is net over-coverage for a benchmark variable.**

**Q3 Does the panel agree that Combinatorial Optimisation is the preferred method for selecting donors?**

**Q4 Are we making best use of administrative data in our plans?**

**Q5 Does the panel have any further suggestions for improving the census coverage adjustment?**

**The 2021 Census Coverage Adjustment Strategy Update**

**External Assurance Panel (October 2019)**

Alison Whitworth, Kirsten Piller, Rajni Sandhu and Andrew Penn

**Introduction.**

The purpose of the census coverage adjustment is to amend the unit level Census database so that it is consistent with the population estimates derived from the Coverage Estimation process, and 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.

The strategy for the 2021 Census is described by Whitworth et al (2018) and was presented at the September 2018 external assurance panel meeting. The paper gives a summary description of the 2011 methods and explains the changes planned for the 2021 Census. The change in methods have been designed to address the practical difficulties that were experienced during implementation of the 2011 methods, and to make best use of the new strategy and outputs from the coverage estimation. It assumes that census population estimates are provided from coverage estimation for a variety of basic demographic characteristics for both households and individuals in local authority (LA) areas, and that these estimates account for patterns of under and over coverage that are typically concentrated within geographical areas and across population subgroups.

Summary of the 2021 adjustment system:

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

1. Derive integer benchmarks for population and household totals by key demographic characteristics that represent the missed households and people within them, using the coverage weights provided by the initial coverage estimation system and rounding.
2. Select donor households using the Combinatorial Optimisation (CO) methods, ensuring the benchmarks in 1 above are maintained as nearly as possible.
3. Place the donor households in an appropriate postcode.

Stage 2. Impute characteristic variables for the persons and households imputed in stage 1.

The justification for simplifying to a two-stage process is that persons missed in counted households will be implicitly corrected through the selection, and placement, of donor households that account for both individual and household characteristic benchmarks. So, for example if a man is missed from a household with a woman and child, instead of imputing a man into an existing two-person household, we impute a three-person household and omit imputation of a two-person household to compensate. The aim is to obtain representative aggregate level population totals rather than an accurate unit level data base. It also reduces complexity in the methodology, and therefore the risk of problems like those experienced in 2011.

In stage 2 imputation of the remaining characteristic variables will be completed using CANadian Census Edit and Imputation System (CANCEIS) methods. While it would be possible to take all the values from a donor record, the values for the remaining characteristics aren’t controlled for by CO so would be less informed.

The key differences to the methods used for the 2011 Census are:

1. Use of a two-stage approach where we impute donors that represent missed households in order to meet the coverage estimation LA benchmarks for both households and persons, and then impute the characteristic variables for these persons and households. In 2011 a three-stage approach was used whereby missed persons were first imputed into counted households.
2. Use of the Combinatorial Optimisation methods to select donor households. The CO methods involve an integer reweighting procedure to find a combination of households from the census database that best fits benchmarks for both households and individuals estimated to have been missed by the Census. In 2011, donor households were selected by comparing a cumulative total of calibrated coverage weights to census counts. CO provides a simpler and more transparent method for selecting (whole) people with the characteristics of non-respondents for use as donors. Further information about the CO methods are available in Appendix A and for a more detailed description see Oguz and Abbott (2016) (Annex A, p49 - 50).
3. Greater use of administrative data for placing donor household into postcodes, potentially providing more robust small area distributions.

The previous strategy paper Whitworth et al, (2018) described some initial work that demonstrated proof of concept for the new methods in 5 estimation areas. 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 methods than when using the 2011 methods. It also described some subsequent analysis showing that the benchmarks were also better met in a two-stage approach compared to a three-stage approach (whereby missed people are first imputed into counted households).

This paper provides more detailed evidence on the implementation of the strategy including performance of the CO method in a two-stage approach and investigation of properties of the CO outputs. We also outline our plans for using administrative data to provide additional evidence to inform the methods.

Section 1 shows the performance of the CO method when using benchmarks for more detailed categorisation of variables at estimation area (EA) level. In 2011 the number of categories for the benchmarks was different across EAs, reflecting the quality of the coverage estimates and the ease with which the calibration could be achieved. By using more detailed and consistent benchmarks with the CO methods we ensure more robust outputs by these key characteristics which are then used as a basis for allocating donors to postcodes and for imputing the other census variables in stage 2.

Section 2 assesses the bias and variance in the CO outputs and provides more evidence on the relative merits of including benchmarks at more detailed categories. The CO procedure starts with a random selection of households and then substitutes households in order to obtain an optimal solution for the benchmarks. Different runs of the CO algorithm would therefore produce a different selection of donors.

Section 3 tests the performance of the CO methods in the two-stage approach under different scenarios for census coverage. It provides evidence on the stability of the CO methods when subjected to different patterns and levels of census non-response (i.e. when the relative proportions of people missed from counted households compared to missed households is substantially greater than in 2011).

Section 4 tests the performance of the CO methods when using LA benchmarks and selecting from a LA donor pool rather than EA area pool. For 2021 it is planned that the coverage estimates will be available at LA level.

Section 5 presents our plans for using administrative data to improve the detail of data used to inform the placement of donors to post codes.

**Section 1 Implementation of the CO method with more detailed variables.**

Earlier investigations of the CO methods (Oguz and Abbott, (2016) and Whitworth et al, (2018)) used the same number of categories for the benchmark variables as used in the 2011 Census methodology. This was so that the performance of the CO method could be compared more directly with the 2011 Census methods. The number of categories used reflected the quality of the coverage estimates (used as benchmarks) and more notably the ease with which calibration could be achieved.

The main issue with the 2011 system was the difficulty calibrating the household weights to both household tenure, and person estimates by age/sex groups. Early runs were failing to select the required number of persons to meet the age/sex by local authority targets, because the household weights were not calibrating correctly. If calibration was not achieved, then all calibrating variables except age-sex group by LA and household tenure were collapsed or dropped. The evaluation of the methodology has also suggested that the three-stage approach may have caused some of the issues.

The methods planned for 2021 Census provides the opportunity to obtain more robust estimates at more detailed levels of disaggregation. North, R., (2019) assesses the performance of the CO method for the same five estimation areas but the number of categories in the benchmark variables is increased. The CO methods are used to find a combination of households from the census database that best fits benchmarks for both households and individuals estimated to have been missed by the census. The fit of the benchmarked variables is assessed using the Total Absolute Error (TAE) which is the sum of the absolute differences between the observed (CO) counts and the expected (benchmark) counts. The selected EAs contain a range of urban and rural areas as well as demographic mix, and some areas faced problems with calibration when using the 2011 system.

The ethnic group, household tenure and hard to count variables were found to be very robust (across all areas considered) achieving a good fit regardless of the way they were categorised. The hard to count index variable was included with the maximum number of categories for each area, typically only two categories (easier or harder to count household). The activity last week variable was typically available as a seven-level variable for each estimation area, but as the error was larger for the age-sex by local authority variable when this was used, the variable was instead used as a four-level variable (full time employed, part time employed, student, other). The variable that had the most notable errors was the household size variable when the number of categories was increased to more than three persons per household. Table 1 below shows the total absolute error (TAE) when different categories for household size are used for the benchmarks for an estimation area within the East Midlands. The donor pool was the census responses for that estimation area, with274,521 households and 624,701 persons.

 **Table 1: Running Combinatorial Optimisation for the whole adjustment process (two stage approach) in estimation area EM05NSDE (East Midlands), with different categorisations of the household size variable.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Benchmarked variable (categories) included in CO process, all but age /sex at estimation area level** | **Average run time per CO run, estimation area/LA benchmarks** | **TAE age sex by LA (average over 100 runs)** | **TAE household size (average over 100 runs)** | **Total TAE across other variables (average over 100 runs)** |
| Key variables: Age-sex by LA (245), tenure (5) | 38 seconds | 30 | NA | 0 |
| Age-sex by LA (245), tenure (5), hard to count index (2), **ethnicity (5), activity last week (4), household size (3)** | 51 seconds | 138 | 1 | 5 |
| Age-sex by LA (245), tenure (5), hard to count index (2), **ethnicity (5), activity last week (4), household size (4)** | 51 seconds | 130 | 143 | 5 |
| Age-sex by LA (245), tenure (5), hard to count index (2), **ethnicity (5), activity last week (4), household size (5)** | 51 seconds | 128 | 287 | 5 |

Note: Total Absolute Error (TAE) is the sum of the absolute differences between the observed counts and the expected (benchmark) counts.

We also observe the time taken to run the procedures and the number of times the same donor is selected. Using the same donor many times could result in infeasible distributions in the output tables, particularly if the selected households have unusual characteristics. For the five estimation areas, and for the different combinations of benchmark variable categories, each CO run took between around 10 and 60 seconds. Also, the CO method did not use the same household as a donor more than approximately 30 times. In contrast, there were cases where the 2011 method used the same household as a donor approximately 300 times. In both methods no restrictions were imposed as to the number of times the same household could be used as a donor, although this is a possibility for 2021.

**Section 2 Variance and bias of the CO outputs and further investigation of household size.**

The CO method takes an initial selection of donors at random and then swaps in donors iteratively in order to improve the fit to the benchmarks. The CO algorithm is run 100 times using the (fixed) 2011 unadjusted census database for a given estimation area to give 100 different selections of donors. The selections are expected to be different with each run. Oguz and Abbott (2016) showed the distribution of the TAEs over 100 runs for the age-sex group by LA variable in all the EAs tested and the household size variable in 3 areas. The size of the error was for CO used in the three-stage approach (so missed persons were first imputed into counted households using the 2011 methods). There were some differences in the distribution between EAs. For the age-sex group by LA variable the Outer London EA showed the greatest range in error and East Midlands the smallest.

North, R (2019) assessed the bias and variance in the CO outputs and further evaluates the benefits of benchmarking to more detailed categories for household size. A two-stage approach is used so donors are selected for missed households without first imputing persons into counted households. The bias for a given category of a given variable is defined as the average of the CO counts over 100 runs minus the true benchmark value from the coverage estimation. A measure of variance is also given in the form of (or approximation for) a confidence interval. For a given category of a given variable, the approximation for the 95% confidence interval is the range within which 95% of the CO counts lie over the 100 runs and is defined by the 2.5 and 97.5 quantiles.

The results obtained for the age sex by local authority, ethnicity, activity last week, household tenure and hard to count index were all found to have little or no bias and low variance for all the areas considered (for the categories considered). Table 2 presents measures of variance and bias in estimation area EM05NSDE for all variables except age-sex by local authority variable which contains 245 categories. However, the bias and variance were low across all the age-sex categories. The other 4 estimation areas give very similar CI widths across the benchmark variables, so most are very small. The only larger CI widths are seen for the household size variable, which again is similar for all EAs, in that there are larger widths for sizes 1 and 5+.

Tables 2 shows the results for the benchmark variables when the household size variable is treated as a five-level benchmark, the advantage of expanding the categories for household size even though the TAE for the variable (as a whole) is greater than when collapsed to 3 categories is highlighted in Table 3. The results are presented with different categorisations of the household size variable whilst keeping the categories for the other variables unchanged. By running the CO process with household size as a three-level variable (1 person, 2 and 3+) the benchmarks for four and five-person households are not being controlled (although outputs will be required for these larger household sizes). Table 3 shows the effect on the four and five-person households when household size is treated as a three-level benchmark in the CO process. Some bias is found in the larger households (3,4, 5+), although the relative biases are small. The results for household size are very similar for all five estimation areas.

**Table 2: Estimates of bias, relative bias and variance for the CO results for estimation area EM05NSDE in the East Midlands, household size as five level benchmark, estimation area level benchmarks and estimation area level donor pool used in CO process**

| Variable | Average CO count over 100 runs | Benchmark value | Bias (= average of (CO count - benchmark value)) | Relative bias | 95% CI for 100 CO counts |
| --- | --- | --- | --- | --- | --- |
| **Tenure** |  |  |  |  |  |
| Owner Occupied | 102774 | 102774 | 0 | 0 | (102773,102775) |
| Owns with mortgage or loan | 97977 | 97977 | 0 | 0 | (97976,97978) |
| Shared ownership | 1226 | 1226 | 0 | 0 | (1225,1227) |
| Renting | 78138 | 78138 | 0 | 0 | (78137,78139) |
| Rent free | 3847 | 3847 | 0 | 0 | (3846,3848) |
|  |  |  |  |  |  |
| **Household size** |  |  |  |  |  |
| 1 | 81824 | 81682 | 142 | 0.0018 | (81814,81838) |
| 2 | 107080 | 107079 | 1 | 0 | (107078,107083) |
| 3 | 44841 | 44841 | 0 | 0 | (44839,44843) |
| 4 | 36181 | 36182 | -1 | 0 | (36175,36182) |
| 5+ | 14036 | 14178 | -142 | -0.0100 | (14023,14050) |
| **Hard to count INDEX (lower = easier to count)** |  |  |  |  |  |
| 1 | 203797 | 203797 | 0 | 0 | (203796,203797) |
| 2 | 80165 | 80165 | 0 | 0 | (80165,80165) |
| **Ethnicity** |  |  |  |  |  |
| White | 631775 | 631775 | 0 | 0 | (631774,631777) |
| Mixed | 7200 | 7200 | 0 | 0 | (7199,7202) |
| Asian | 8161 | 8161 | 0 | 0 | (8160,8163) |
| Black | 2138 | 2138 | 0 | 0 | (2137,2140) |
| Other | 808 | 808 | 0 | 0 | (807,810) |
| **Activity last week** |  |  |  |  |  |
| Employed | 316280 | 316280 | 0 | 0 | (316279,316282) |
| Unemployed | 28763 | 28763 | 0 | 0 | (28762,28766) |
| Student | 17446 | 17446 | 0 | 0 | (17445,17448) |
| Other | 287593 | 287593 | 0 | 0 | (287592,287595) |

Note: The results for the age-sex by local authority variable are not shown for this estimation area as the variable contains 245 categories, which would lead to overly large tables. The bias and variance were low across all categories.

Tables 2 shows the results for the benchmark variables when the household size variable is treated as a five-level benchmark. Overall the best results are obtained when household size is treated as a five-level benchmark since it yielded the lowest bias and variance for the large households and allows control for the larger household size categories. Although it yielded some bias being introduced in the one-person households, it was not substantial. The estimates for the other benchmark variables are unbiased with low variance regardless of the way in which household size is categorised.

The analysis demonstrated that overall the CO methods perform better than the 2011 when benchmarking to more detailed categories although there may be some trade-off between variables in order to achieve optimal quality estimates for key outputs. The best solution is likely to be specific to each variable.

**Table 3: Estimates of bias, relative bias and variance for all household sizes when household size is treated as a three-level benchmark in the CO process, CO results for estimation area EM05NSDE in the East Midlands, estimation area level donor pool used.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Average CO count over 100 runs | Benchmark value | Bias (= average of (CO count - benchmark value)) | Relative bias | 95% CI for 100 CO counts |
| **Household size (4 categories)** |  |  |  |  |  |
| 1 | 81862 | 81862 | 0 | 0 | (81681, 81683) |
| 2 | 107079 | 107079 | 0 | 0 | (107078,107080) |
| 3 | 45427 | 44841 | 586 | 0.0131 | (45368,45479) |
| 4+ | 49774 | 50360 | -586 | -0.0116 | (49722,49833) |
| **Household size (5 categories)** |  |  |  |  |  |
| 1 | 81862 | 81862 | 0 | 0 | (81681, 81683) |
| 2 | 107079 | 107079 | 0 | 0 | (107078,107080) |
| 3 | 45427 | 44841 | 586 | 0.0131 | (45368,45479) |
| 4 | 35704 | 36182 | -478 | -0.0132 | (35624,35783) |
| 5+ | 14070 | 14178 | -108 | -0.0076 | (14030,14115) |

Note: Estimates of bias, relative bias and variance for other variables (when household size is treated as a three-level benchmark in CO process) are provided in Appendix B.

**Section 3 The two-stage approach under different scenarios for census coverage.**

The justification for simplifying to a two-stage process is that persons missed in counted households will be implicitly corrected through the selection, and placement of donor households that account for both individual and household characteristic benchmarks. So, for example if a man is missed from a household with a woman and child, instead of imputing a man into an existing two-person household we impute a three-person household and omit imputation of a two-person household to compensate. The aim is to obtain representative aggregate level population totals rather than an accurate unit level data base. It also reduces complexity in the methodology, and therefore the risk of problems like those experienced in 2011.

In this section the performance of the CO methods in the two-stage approach is tested under a different scenario for census coverage(e.g. when the proportions of people missed from counted households compared to missed households is substantially greater than in 2011). The objective was to test the limits of the CO method in compensating for people missed from counted households in the two-stage approach. (This is work in progress, so a more detailed paper is not yet available for reference.)

For this analysis persons were removed from counted households in the unadjusted census database thus creating different coverage scenarios. Counted households were not removed, because the aim was to increase the relative proportion of persons missed from counted as opposed to missed households. We believed this would be the most challenging scenario for the 2021 strategy. Further work will test coverage patterns whereby numbers of missed households are increased as this is the most likely scenario with lower census response rates. The benchmarks were kept unchanged on the basis that the coverage estimation would have accurately accounted for the different coverage scenarios (although in practice this may be overly optimistic particularly for household size). In the scenario tested here the unadjusted census database would contain a greater number of smaller households due to the removal of extra persons. The numbers of persons imputed into counted households in 2011 were used as a basis for deciding how many extra persons to remove from counted households.

Three of the 5 EAs used in the previous analyses were selected for this initial assessment and comparisons made with the previous results. A random selection of persons was removed from the donor pool (unadjusted 2011 Census data) for each EA; for the first 2 EAs (EM05NSDE (East Midlands, North and South Derbyshire) and SW08MIDE (South West, Mid and East Devon and Somerset)) the removals resulted in twice the number of persons missed from counted households compared to 2011. The final estimation area (OL13WALT (outer London, Waltham Forest)) resulted in 1.5 times the number missing, as removing a greater proportion created a greater number of two person households than was evident in the coverage estimates. The random selection of persons to be removed was partly informed by estimated probabilities of being missed from a counted household, obtained from the 2011 adjustment system. In other words, persons were selected from respondents who had a comparatively high estimated probability of being missed from a counted household. To date this assessment has only been run with one random selection of persons being removed from each estimation area’s donor pool, although the results will likely vary for different selections. In our future work we plan to extend this initial analysis by removing different selections of persons in simulations.

**Table 4: Estimates of bias, relative bias and variance for the CO results for estimation area EM05NSDE in the East Midlands, household size as five level benchmark, estimation area level benchmarks and estimation area level donor pool used in CO process, *with a random selection of persons removed from counted households***

| Variable | Average CO count over 100 runs | Benchmark value | Bias (= average of (CO count - benchmark value)) | Relative bias | 95% CI for 100 CO counts |
| --- | --- | --- | --- | --- | --- |
| **Tenure** |  |  |  |  |  |
| Owner Occupied | 102774 | 102774 | 0 | 0 | (102772, 102776) |
| Owns with mortgage or loan | 97977 | 97977 | 0 | 0 | (97976, 97978) |
| Shared ownership | 1226 | 1226 | 0 | 0 | (1224, 1227) |
| Renting | 78138 | 78138 | 0 | 0 | (78137, 78139) |
| Rent free | 3847 | 3847 | 0 | 0 | (3846, 3848) |
| **Household size** |  |  |  |  |  |
| 1 | 81734 | 81682 | 97 | 0.0012 | (81747, 81806) |
| 2 | 107080 | 107079 | 2 | 0 | (107078, 107089) |
| 3 | 44941 | 44841 | 13 | 0.0001 | (44841, 44889) |
| 4 | 36182 | 36182 | -1 | 0 | (36179, 36183) |
| 5+ | 14025 | 14178 | -111 | -0.0078 | (14039, 14092) |
| **Hard to count INDEX (lower = easier to count)** |  |  |  |  |  |
| 1 | 203797 | 203797 | 0 | 0 | (203797,203798) |
| 2 | 80165 | 80165 | 0 | 0 | (80164,80165) |
| **Ethnicity** |  |  |  |  |  |
| White | 631775 | 631775 | 1 | 0 | (631774,631778) |
| Mixed | 7200 | 7200 | 0 | 0 | (7199,7202) |
| Asian | 8161 | 8161 | 0 | 0 | (8160,8164) |
| Black | 2138 | 2138 | 0 | 0 | (2137,2141) |
| Other | 808 | 808 | 0 | 0 | (807,810) |
| **Activity last week** |  |  |  |  |  |
| Employed | 316280 | 316280 | 1 | 0 | (316279,316283) |
| Unemployed | 28763 | 28763 | 1 | 0 | (28762,28767) |
| Student | 17446 | 17446 | 1 | 0.0001 | (17446,17449) |
| Other | 287593 | 287593 | 1 | 0 | (287592,287596) |

Table 4 provides measures of the bias and variance around the 100 CO runs completed for the East Midlands EA and can be compared with Table 2 above. The bias in the CO outputs for the benchmarked variables remains low across most categories and is only slightly higher compared to Table 2 for household size. The South West and Outer London EAs gave very similar errors across these categories and the age-sex by LA categories, compared to the original work in section 2 where the 2011 coverage patterns are unchanged.

However, the East Midlands EA had larger error for the age-sex by LA variable, this variable is not available in the table as it has 245 categories. For the assessment of CO with the original unadjusted census data the TAE (based on biases calculated as in below table) was around 125 and after removing a random selection of persons from the unadjusted data it was around 1125. Again, this is spread across the 245 categories although not always evenly.

Although the increase in TAE for the East Midlands area is substantial, it is similar to that from the 2011 methods. The greater risk associated with a change in coverage patterns is demonstrated by OL13WALT (Outer London, Waltham Forest).

The 2021 adjustment strategy assumes that there is a net under coverage in the unadjusted census tables and currently does not cater for a situation where the total population for a benchmarked variable is greater in the unadjusted census database than in the coverage estimate (benchmark). The CO method adds households to meet both person and household level benchmarks but does not remove households. As described above when the number of persons missed from counted households in the estimation area OL13WALT (outer London, Waltham Forest was increased by more than 50% of the 2011 figure, the number of households of size 2 persons were greater than the coverage estimation benchmarks for this variable. Analysis of another EA showed that the number of households of both size 1 and 2 were greater than the coverage estimation benchmark. Although in practice it may be unlikely, we will need to investigate a contingency strategy for a scenario that results in a net over-coverage in the individual benchmark variables.

**Section 4 Performance of the CO methods when using LA benchmarks and selecting from a LA donor pool rather than EA areas.**

Three LAs were selected from the same EAs for this initial analysis. The Inner and Outer London EAs that have been used in the CO research are both equivalent to one LA but are large enough in size to be considered an EA for 2011 Census. As the objective of the analysis was to test the CO within a smaller geographical area than previously, one LA was selected from each of the 3 EAs that contained more than one LA. We used the adjusted census estimates to obtain LA benchmarks for all of the variables previously used as at the time 2011 coverage estimates were only available at LA level for the age-sex variable. This means that we cannot make comparisons to the adjusted 2011 Census estimates, but the intention is to later use simulated coverage estimates to test the methods further when run for LAs and using additional variables.

We found that there was minimal bias across the age-sex, tenure, hard to count index, activity last week and ethnicity benchmark variables, in fact for the one LAs selected in the Yorkshire and Humberside estimation area there was no bias across all 6 benchmark variables (including household size). Bias was found for the 2 LAs in the East Midlands and South West EAs, although for the latter the bias was extremely small.

**Table 5: Estimates of bias and variance for the CO results for LA 17UH in EM05NSDE in the East Midlands, household size as five level benchmark, LA level benchmarks and LA level donor pool used in CO process**

|  |  |  |
| --- | --- | --- |
| Variable | Bias (= average of (CO count - benchmark value)) | 95% CI width for 100 CO counts |
| **Tenure** |  |  |
| Owner Occupied | 0 | 0 |
| Owns with mortgage or loan | 0 | 0 |
| Shared ownership | 0 | 1 |
| Renting | 0 | 0 |
| Rent free | 0 | 0 |
| **Household size** |  |  |
| 1 | -261 | 29 |
| 2 | 0 | 3 |
| 3 | 0 | 2 |
| 4 | 13 | 39 |
| 5+ | 248 | 33 |
| **Hard to count INDEX (lower = easier to count)** |  |  |
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| **Ethnicity** |  |  |
| White | 0 | 1 |
| Mixed | 0 | 1 |
| Asian | 0 | 1 |
| Black | 0 | 1 |
| Other | 0 | 0 |
| **Activity last week** |  |  |
| Employed | 0 | 1 |
| Unemployed | 0 | 1 |
| Student | 0 | 1 |
| Other | 0 | 1 |

Note: The local authority level constraints and any other columns using them cannot be included in this table, as they are based on other census data and are not coverage estimates.

Table 5 shows the error for the LA in the East Midlands, which has bias for household size and larger relative bias as the target population counts for LAs are smaller. We expect to have better evidence of how well the methods work when implemented for LAs with future work using simulated coverage estimates, but this early work has indicated that CO can optimise well with smaller donor pools and benchmarks.

**Section 5 Placement of donor households and use of administrative data.**

In this section we present plans for using administrative data to better inform the placement of donors in post codes. Once the donor households have been selected, we look to other data to decide where the households should be placed. The census (“dummy form”) data provide us with indicators of potentially missed households, and administrative data may provide additional information to support this or to corroborate what is provided by the census.

The methods used to place household donors for the 2011 Census were carried out by local authority for each estimation area, and the donors were placed in two steps. Donor households were first placed at the location of a census dummy form (i.e. an address where the enumerator was unable to obtain a response). A simple scoring method was used that compared the characteristics of the donor households to the characteristics of the available addresses and their output areas. When the dummy forms were used up within a local authority, or there were no appropriate dummy forms left, donor households were placed randomly within the donor’s OA.

The simple scoring method aims to place donor households in to the locations of similar dummy forms or empty households. The accommodation type, household structure, tenure (either at address or at output area level) and lower level super output area (LSOA) of each donor were compared to the dummy forms. The score is increased where these characteristics were the same. Dummy forms that provided the largest score and smallest distance for the donor were selected as the location of placement for each donor.

Of the 348 local authorities, 162 went on to place donor households randomly after first trying to place the donors in dummy forms/empty households. Of these LAs, 107 moved on to random placement as they had filled all the available dummy forms and had donors still to place. The other 55 LAs did not use all dummy forms as there were very low scores between the remaining dummy forms and donors.

**Table 6: Placement of donors within LAs. Use of dummy forms by random placement in 2011 census adjustment.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Used all dummy forms for placement** | **Didn’t use all dummy forms for placement** | **Total** |
| **Did random placement** | 107 | 55 | 162 |
| **Didn’t do random placement** | 0 | 186 | 186 |
| **Total** | 107 | 241 | 348 |

Administrative data linked to the 2021 Census records could potentially improve the placement of the donor households and be used in the following ways:

### Additional variable information at unit or area level

For records where there is a link between the census and an administrative data source (either dummy form records or census responses) we could make use of additional or alternative variables from the administrative data. So, for the dummy form records in their unit level comparisons to the donor households. For example, if dummy forms were linked to council tax data with a variable indicating a council tax discount for a one-person household, donors with a household size of one could have an increased score for these dummy form records. Or, household structure, which was used at output area (OA) level for the dummy forms in 2011, could be derived from the “Administrative based population estimates” (ABPE) to allow a unit level comparison where there wasn’t one before (assuming the ABPE could be available in time).

Another option may be to create OA level indicators from the additional variables to inform placement of a donor household in a dummy form or postcode, rather than compare at the unit level. Or, use OA level distributions rather than indicators, so there would be information about how prevalent a variable’s category (e.g. proportion of households/properties with the accommodation type detached) is in an area.

Although it is likely that not all census records and dummy forms will be successfully linked to the specific administrative source via the UPRN, additional information (from administrative data) for the records that are successfully linked would still (potentially) improve the placement. Obtaining the best match of the donors to postcodes using whatever evidence is available to inform this, can be undertaken on a case by case basis. OA (as well as unit) level indicator variables may also help; however a judgement will need to be made as to whether the census responses provide more complete and higher quality OA level information even though they exclude census non-respondents.

### Validation

If the administrative data cannot be used directly, it could potentially be used at the end of the adjustment to validate the number or characteristics of households that have been placed by geographical area, with consideration of the possible difference in definitions (i.e. household vs. address) and how recent activity/interaction with a service was at an address.

Table 7 shows the information available in different administrative data sources and how it might help to inform the placement of donors.

**Table 7 Potential variable information available by administrative data source for use within the census adjustment**

|  |  |  |
| --- | --- | --- |
| **Source** | **Variable** | **Comparison with donor** |
| VOA | Number of bedrooms | Household size? |
| VOA | Property type | Accommodation type with more detailed categories than census |
| Council tax | Flag to indicate exemption from paying council tax for student household | Activity last week |
| Council tax | Flag to indicate discount in council tax payment for single person household | Household size |
| Council tax | Flag to indicate discount in council tax payment for empty or second home | flag for not to be used as a space. |
| ABPE | Household composition | Household structure |

Note: The value of the linked data at the record level would depend on the quality of the linkage and the variables.

We also considered using administrative data as a direct indicator of a missed household (in a similar way to the way the dummy forms are used). One option, for example, is to use administrative data to provide unit level information about households that have potentially been missed. We considered treating VOA or ABPE records that didn’t link to census respondent records or dummy form records as an indicator for a missed household in a similar way to the dummy forms. The records could be compared to donors, so if it shows evidence of recent activity and enough information to say they are similar, then the postcode of the administrative data record could be used for the donor.

We decided against this approach however, because administrative data will be used to improve and validate the address register, so we would expect a census response or dummy form for valid entries. Entries on the administrative data that cannot be linked to census data or dummy forms may be for properties that no longer exist, be duplicate addresses or the UPRN may have been incorrectly assigned in the VOA or census data. The benefit above using a random placement is therefore unclear.

**Summary and conclusions**

Previous research (based on 5 EAs) has shown proof of concept that the selected donors when using the CO method better meet the benchmarks than the 2011 method (Oguz and Abbott , (2016)), and also that both household and person level benchmarks are successfully met using a two-stage approach whereby only missed households are imputed without first imputing missed persons from counted households (Whitworth et al, (2018)). This paper has shown more detailed evidence on the properties of the CO outputs and implementation of the two-stage approach.

Section 1 discussed the performance of the CO methods when using benchmarks for more detailed categorisation of variables at EA level. For most variables (ethnic group, household tenure and hard to count variables) a good fit was achieved with more detailed categories. For the household size variable and activity last week variable though, there was greater error when more detailed categories were used. The best categorisation detail will be specific to each variable: research has shown that for the household size variable benchmarking to a more detailed categories is preferable (Section 2) whilst for the activity last week variable, it is preferable to use the collapsed categories in order to avoid greater error in the age-sex by local authority variable which is the key output.

Further work will test the performance of the CO method when additional variables (e.g. marital status) are included in the benchmarks. The advantage of using more detailed benchmarks with the CO methods is to ensure more robust outputs for use in allocating donors to postcodes and for imputing the other census variables in stage 2, and ultimately more robust outputs at small area level.

Overall the CO outputs demonstrated relatively low bias for most variables and very low variance in (Section 2). Section 3 showed that in a different scenario whereby the number of people missed from counted households was higher than in the 2011 Census (but the number of missed households remained unchanged), measures of the bias and variance for some of the EAs tested (East Midlands EA The South West and Outer London EAs) remained low across most categories or gave very similar errors. The East Midlands EA showed considerably larger error under this scenario for the age-sex by LA variable, but it was still not very much larger than the magnitude of error evident in 2011 (with 2011 coverage patterns).

Section 3 also showed that in this alternative coverage scenario for some LAs there were more small households than evident in the estimated coverage benchmarks. The adjustment strategy deals only with net under coverage so a contingency strategy will need to be developed for the 2021 Census. In our future work we plan to extend this initial analysis to other areas and by removing different selections of persons in simulations. We will also test the methods in scenarios whereby there are increased numbers of missed households.

Section 4 tests the performance of the CO methods when using LA benchmarks and selecting from a LA donor pool rather than EA area pool. This early work has indicated that CO can optimise well with smaller donor pools and benchmarks. Of the 5 LAs tested we found that there was minimal bias across the age-sex, tenure, hard to count index, activity last week and ethnicity benchmark variables, however some bias was found the LAs in the East Midlands for the household size variable.

Finally, in section 5 we outlined our plans for using administrative data for the census adjustment. The most valuable use is expected to be in providing additional evidence with which to place donors into appropriate postcodes. For records where there is a link between the census (or dummy forms) and an administrative data source, we could make use of additional variables from the administrative data to better inform the pairing of dummy form records to the donor households. For example, using council tax data with a variable indicating discount for a one-person household. For efficiency this should be carried out as data preparation before the adjustment processing begins and is dependent on successful quality assurance and linking of census data. We plan to specify the enhancements to the information provided with dummy form data within our data requirement specification.

**Summary of future work**

1. Further testing of the CO method to better understand the properties of the CO outputs. This will include:
* Using additional benchmarks for CO to assess whether this improves the imputation for smaller geographical areas.
* The processing times for running the adjustment with more benchmarks. To some extent longer running times could be trader for improvements in the quality of the outputs.
* Testing the methods in additional EAs and LAs and the limits of the method
* Further analysis different coverage scenarios, including a greater number of missed households.
* The methods to impute individuals missed from Communal Establishments (CEs).
1. The variability of estimated variables: To date emphasis has been on using a fixed unadjusted census database for a given area then running the CO 100 times to give 100 different selections of donors. The plan is to develop an option to run multiple times and choose a central “minimum error” completion to be used as donors. This will be extended by simulating multiple unadjusted census databases from the true census database and running the CO method for each unadjusted census database. This would allow an assessment of the variability and bias in different unadjusted census databases for a given area as well as the random component of the CO procedure. In the longer term this could be added with the variance estimation of whole estimation procedures to create measures of uncertainty.

**References**

North, R., (2019), “Combinatorial Optimisation for Census adjustment in 2021”, Internal working paper.

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).

**Appendix A**

**Combinatorial Optimisation (CO) method**

CO is an integer programming method which involves finding the best combination (solution) from a finite set of combinations for a problem (Voas and Williamson, 2000). In the context of the census coverage adjustment, CO involves the selection of a combination of households from the unadjusted census database that best fits the estimated benchmarks. In this context the benchmarks (or constraints) are totals of the individual and household variables in the local authorities/estimation areas from the coverage estimation process (i.e. of those who were missed from the census enumeration). When running the CO process, the benchmarks are specified as the shortfall between the unadjusted census database count and the corresponding benchmark total.

The CO approach is essentially an integer re-weighting exercise where most of the households in the census database are assigned zero weights and positive integer weights are assigned to a combination of households which satisfies the required constraints (imputation totals) for both households and individuals. Therefore, this approach can be considered as an alternative method for imputing households (and individuals within the households) estimated to have been missed by the census (Oguz and Abbott, 2016).

The research has used CO with the simulated annealing, which has been found to perform best with CO. Simulated annealing determines whether a swapped in household is accepted. Household record swaps which adversely affect the fit might be accepted in order to avoid getting trapped with a sub-optimal selection of households.

NB Oguz and Abbott, (2016) also provide more detailed description of the methods.

**Appendix B: Further investigations of the use of the CO for the census adjustment process**

**Table 1: Estimates of bias, relative bias and variance for the CO results for estimation area EM05NSDE in the East Midlands, household size as three level benchmark, estimation area level benchmarks and estimation area level donor pool used in CO process.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Average CO count over 100 runs | Benchmark value | Bias (= average of (CO count - benchmark value)) | Relative bias | 95% CI for 100 CO counts |
| **Tenure** |  |  |  |  |  |
| Owner Occupied | 102774 | 102774 | 0 | 0 | (102773,102775) |
| Owns with mortgage or loan | 97977 | 97977 | 0 | 0 | (97976,97978) |
| Shared ownership | 1226 | 1226 | 0 | 0 | (1224,1227) |
| Renting | 78138 | 78138 | 0 | 0 | (78137,78139) |
| Rent free | 3847 | 3847 | 0 | 0 | (3846,3848) |
| **Household size** |  |  |  |  |  |
| 1 | 81862 | 81862 | 0 | 0 | (81681, 81683) |
| 2 | 107079 | 107079 | 0 | 0 | (107078,107080) |
| 3+ | 95201 | 95201 | 0 | 0 | (95200,95202) |
| **Hard to count INDEX (lower = easier to count)** |  |  |  |  |  |
| 1 | 203797 | 203797 | 0 | 0 | (203796,203797) |
| 2 | 80165 | 80165 | 0 | 0 | (80165,80166) |
| **Ethnicity** |  |  |  |  |  |
| White | 631775 | 631775 | 0 | 0 | (631774,631777) |
| Mixed | 7200 | 7200 | 0 | 0 | (7198,7202) |
| Asian | 8161 | 8161 | 0 | 0 | (8159,8163) |
| Black | 2138 | 2138 | 0 | 0 | (2136,2139) |
| Other | 808 | 808 | 0 | 0 | (807,809) |
| **Activity last week** |  |  |  |  |  |
| Employed | 316280 | 316280 | 0 | 0 | (316278,316281) |
| Unemployed | 28763 | 28763 | 0 | 0 | (28761,28765) |
| Student | 17446 | 17446 | 0 | 0 | (17444,17447) |
| Other | 287593 | 287593 | 0 | 0 | (287591,287595) |