**The 2021 Census Coverage Adjustment Strategy**

**Paper for External Review (September 2018)**

Alison Whitworth, Christine Sexton and Rob North

1. **Introduction**

The purpose of the census coverage adjustment method 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.

In 2011 the census coverage estimation used dual system, and ratio estimators to produce census-based population estimates for a number of key household and individual demographic characteristic variables at Local Authority (LA) and Estimation Area (EA) levels (ONS, 2013). The dual system estimator used the outcome of matching the census to the census Coverage Survey (CCS), and estimates were produced for around one hundred EAs (which consist of one or more LAs). During the Census Adjustment stage, households and individuals estimated to have been missed from the census were imputed onto the census database across the whole estimation area, so that the household and individual level database was consistent with the population estimates already produced.

The methods were designed to accommodate the “batched” delivery of raw census data as the paper questionnaires were processed following the 2011 Census. In 2021 the online census enumeration will change the delivery of raw census data and although there are challenges and risks associated with this, it potentially enables more efficient and improved methods for the coverage estimation and adjustment.

This paper sets out the strategy for the 2021 Census Coverage Adjustment; it assumes that the census population estimates are available for a variety of basic demographic characteristics for both households and individuals in Local Authority areas, and that these estimates account for patterns of under and over coverage that are typically concentrated within geographical areas and across population subgroups. The new adjustment strategy addresses two main factors; the practical difficulties that were experienced during implementation of the 2011 methods and secondly the new strategy and outputs from the initial coverage estimation.

1. **Summary of 2011 Adjustment Strategy**

The 2011 adjustment strategy detailed by Brown et al. (2011) was based on the 2001 methods developed by Steele et al. (2002). In summary it consisted of three main stages:

1. Imputation of persons missed from counted households,
2. Imputation of missed households (and persons within them), and
3. Imputation of characteristic variables for the persons and households imputed in stage 1 and 2.

The same basic steps were used for imputing missed persons and then households (i.e. the steps were repeated for both stages 1 and 2):

1. Derive census weights which represent the probability of a person or household being missed from the census.

These weights were derived using the matched Census-Census Coverage Survey data to model persons missed in counted households (stage 1) and missed households (stage 2). Multinomial and binomial logistic regression models were used respectively to model the probability of being missed in the census enumeration against individual and household level characteristics that were associated with non-response (including the hard to count (HtC) index). The appropriate coverage weights were calculated from these models.

1. Calibrate the weights to coverage estimation benchmarks (i.e. higher-level estimates at EA and LA levels).

The coverage weights derived in step i) above were calibrated to the census population estimates derived from the estimation system. The calibration was attempted by applying a raking ratio estimator where the weights were iteratively scaled to estimates for each variable in turn until a convergence criterion was achieved. The benchmarks included 5-year age, sex group totals by LAs and other key characteristics at EA level (see Appendix 1).

1. Selection of persons/households to be imputed.

Cumulative (calibrated) weights were calculated and compared with the sum of persons or households for each LA at a time. When the weighted count became 1 or greater (compared to the unweighted count) this indicated that a person / household needed to be imputed, so the response was copied as a donor. The data were sorted by age/ sex group, calibrated weight, Output Area code and postcode prior to this step to ensure a representative distribution of donors across these variables (at least to some extent).

1. Placing missed persons/ households (imputation).

The donor persons from steps i) to iii) (stage 1) were placed into households that were similar to that of the donor’s household had they had been missed where possible. For example, if a donor is a man with a wife and child then he would be placed into a household with a woman and child. Where it was not possible to match the donor and recipient households the criteria were progressively relaxed.

For the donor households from steps i) to iii) (stage 2) the household with its given characteristics was placed in the location where a suitable dummy form or “placeholder” was identified with the same characteristics. If a suitable form did not exist the donor household was placed in a random postcode within the donors output area. Dummy forms were completed by the enumerator where a residence, household or accommodation was identified they but they were unable to achieve a response. They include some very basic information about the type of accommodation.

The imputed persons and households, together with some key characteristic variables, formed “Skeleton records”. In stage 3, imputation of the remaining characteristic variables was completed using CANadian Census Edit and Imputation System (CANCEIS) methods and software. Further information on CANCEIS is available in the Edit and Imputation Process report (ONS, 2012).

As described by ONS (2013), the adjustment methods were successful and worked reasonably well for the 2011 Census outputs, however there were some practical difficulties during implementation. In particular, the calibration did not always work well when constraining to both individual and household variables simultaneously. The solution was to cut the number of constraints when the optimal point in the calibration was achieved, then final convergence was obtained based only on a smaller selection of key variables. This was successful in that constraints for the most important variables were met well (e.g. age, sex totals and household tenure) but less so in that the constraints for household size and household ethnicity were not always met. Furthermore, the methodology was developed late due to other priorities and lack of resource. This meant that testing the methodology in full was not undertaken, and this resulted in some of the issues with calibration described above. Evaluation studies also found that the coverage models used were not always well specified and did not always fit well, due to the relatively small sample sizes for some groups in some EAs.

1. **The 2021 Online Census**

For 2021 there will be an online census of all households and communal establishments in England and Wales with special care taken to support those who are unable to complete the census online. The objective is to maximize response and minimize variability in response rates by optimising the census data collection.

The online collection means that gains in accuracy and processing speed could potentially be achieved in the coverage estimation strategy as suggested by Ross and Abbott (2015), for example the census data for the entire country are going to be available for estimation quicker than previously, and therefore batching into estimation areas will not be necessary. The strategy developed by Racinskij (2018) is that the Coverage Estimation can be carried out either at national or at regional levels rather than for EAs as in 2011, and census coverage can be estimated for a larger set of characteristics. The plan is that logistic or mixed effects logistic regression models will be used to obtain non-response weights for individuals / households with given characteristics and the weights can then be applied to each census return with the same set of characteristics. The sum of the weighted census observations will provide estimates for the population subgroup of interest.

As well as higher overall accuracy both at national and subnational levels the models may potentially provide weights by a wider set of variables (other than the effects of age, sex, HtC, geography and other covariates used). Higher level population estimates for variables such as ethnicity, activity last week, tenure, household relationships etc. would then be available by Local Authority from a single estimation process.

1. **2021 adjustment strategy**

The strategy for 2021 Census adjustment aligns the methods with those used for the higher-level coverage estimation, and addresses the implementation difficulties encountered with the 2011 adjustment system. Specifically, it will make use of Local Authority census population estimates for the wider range of key demographic characteristics produced by the estimation system, as constraints, and uses the combinatorial optimisation method (a synthetic micro simulation technique) in place of the calibration methods to obtain a unit level database that meets these high-level benchmarks, as explored by Oguz and Abbott (2016). The adjustment will be made on a net basis, i.e. the assumption is that under-coverage will always be larger than over-coverage and so the methodology will only create new records and will not remove any existing records on the census database. Should this assumption be violated, and over-coverage is significant, then a separate methodology for dealing with over-coverage would need to be developed. At present, there is no requirement or plans to develop such an approach.

It is proposed that the adjustment system is simplified from a three stage to a two-stage approach for 2021, to:

1. Impute missed households (and persons within them)
2. Impute characteristic variables for the persons and households imputed in stage 1,

and has the following steps at stage 1:

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
2. Select donor households using the combinatorial optimisation methods, ensuring the benchmarks in 1 above are maintained
3. Place the donor households in an appropriate output area

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.

The combinatorial optimisation method replaces the following steps in the 2011 system:

1. Calibrate weights to coverage estimation benchmarks (at EA and LA levels)
2. Selection of persons/households to be imputed

In stage 2 imputation of the remaining characteristic variables will be completed using CANCEIS methods and software.

1. **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).

1. **Research undertaken to data to implement CO methods**

Oguz and Abbott (2015) describe the initial research to implement the Combinatorial Optimisation method and compare results with those from the 2011 methods. This focused on investigating how well the final adjusted census database reflects the estimated benchmark variables when using the different methods and compares the performance of the CO to the 2011 method. The CO method was implemented to impute wholly missed households within the 2011 adjustment system for five Estimation Areas (so persons missed from counted households had already been imputed using the 2011 methods).

The research showed that the CO approach produced better overall, and distributional fit than the 2011 method for most of the benchmarked variables. Using the Total Absolute Error as a measure of performance for each category of a given variable (i.e. the sum of the absolute differences between the observed counts and the expected (benchmark) counts), it was found that running CO for the wholly missing households leads to an improvement in performance for most variables compared to the 2011 system in all five estimation areas. However, it performed less well for the household size variable in one of the estimation areas (Inner London). This may have reflected the small donor pool of the characteristics of this area. The values of benchmarked joint distribution of the variables were also evaluated and were found to be similar to the distributions of the 2011 adjustment system.

The GSS Methodology Advisory Committee agreed that the results were encouraging and demonstrated the feasibility of this method for use in the adjustment system and recommended that the approach should be pursued. The CO method has a considerable performance advantage in terms of its simplicity and computational efficiency over the 2011 method.

Subsequent work for this paper has investigated use of the CO method to impute missed households and persons within them (but without first imputing missed people to counted households using the 2011 system). Again, the investigation used the Total Absolute Error (TAE) as a measure of performance for each category of a given variable, comparing the performance of CO within a two stage and three stage approach against the 2011 method.

Whilst running CO for the wholly missing households only (i.e. the first investigation) leads to an improvement in TAE for all variables compared to the 2011 system except for household size in inner London, running CO for the whole adjustment process leads to an improvement in TAE for all variables compared to the 2011 system in all five estimation areas. Within the two stage approach results showed the household size benchmark being perfectly or nearly perfectly met in all five estimation areas unlike the three-stage approach. An explanation for this could be that there are more one-person households (if missed persons are not imputed into counted households) and CO more easily finds the combinations required with smaller households than with larger households. However, in some estimation areas, the TAE is higher for the age sex by local authority variable when CO is run for the whole adjustment rather than for the missing households only (Table 1 below). The age sex by local authority variable is the most important variable (being the primary output) but has many categories so the difference in TAE is small (see Appendix 2 for the detailed results of all five estimation areas).

**Table 1: East Midlands Estimation Area**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (number of categories) | Total Absolute Error | | |
| CO for wholly missing households (average of 100 runs) | **CO for whole adjustment process (average over 100 runs)** | 2011 system |
| Age-Sex Group by LA (245) | 67 | **118** | 652 |
| Ethnicity (3) | 1 | **1** | 5132 |
| Activity Last Week (2) | 1 | **1** | 1130 |
| Household size (3) | 277 | **1** | 1106 |
| Hard to Count Index (2) | 0 | **0** | 396 |

For this initial work on the CO method the benchmark variables used were the same as those used in the 2011 Census methods in order to ensure fair comparison. However, in 2011 some variables had been collapsed to assist the calibration process. Work as therefore been undertaken to run the CO method for the same 5 estimation areas when the number of categories in the benchmark variables is increased. Again the methods are evaluated using the total Absolute Error (as compared to the benchmark values).

The research to date has also investigated whether the Item imputation (to complete other un-benchmarked characteristic variables of the imputed persons and households) at stage 2 in the 2021 strategy could be replaced with the CO methods (so the values of the selected donors would be retained rather than new values imputed). However, the CANCEIS methods are designed to ensure that the values of the imputed variables reflect average empirical distributions and that unusual characteristics are not propagated within the processing to cause unrealistic distributions (ONS, 2011). For example, the modularised nature of the imputation strategy in 2011 (i.e. variables split into demographics, culture, health and labour market), maximised potential donors at each stage and reduced the likelihood of all records with the same distribution of skeleton variables having exactly the same donor for all target variables (99% of observed records are unique across all variables). CANCEIS also includes a mechanism for obtaining low level geographic consistency, specifically it favours records as donors that are nearest neighbors both in the terms of characteristics *and* geographic location. It would introduce unnecessary risk to change this stage of the 2011 method unless further empirical research proves that the complexity of the CANCEIS methods is unnecessary for census adjustment in 2021*.*

1. **Next steps**

The work to date has focused on comparing outputs using the CO methods with those using the 2011 methods, and initial testing of a two-stage approach. It has provided proof of concept; the CO method obtained an integer donor pool that better met the constraints for characteristic variables. It also has shown potential in providing benefits in terms of efficiency and processing speed.

Further work will focus on three main components:

1. Assessment of the properties of outputs achieved using the CO methods. A simulation study will be used to investigate:
   * The performance of the method as the number of constraints increases and how best to deploy the benchmarks
   * The limits of the method, i.e. for which census coverage patterns the CO does not produce globally optimal dataset
   * The variability of estimated variables, including lower level geographical areas
   * The processing times for running the adjustment by regions and with more constraints
2. Investigation of the wider implications of the new adjustment strategy, specifically how it aligns with;
   * The higher-level coverage estimation strategy; defining requirements for coverage weights for the new census adjustment strategy
   * The methods to place the donor households; whether the 2011 methods meet the requirements of the new 2021 adjustment strategy and whether any improvements can be made
   * The final item imputation; a short empirical assessment of the impact of using CANCEIS methods against retaining the values of the donor households imputed using the CO approach.
   * The methods to impute individuals missed from Communal Establishments (CEs). In 2011 the missed residents were imputed into CEs by selecting donors from the existing residents within the CEs. There was not a mechanism to impute whole CEs as it was assumed that the census would enumerate them accurately.
3. Measurement of the quality of the census database based on simulations, including;
   * The variability of estimated characteristics for (a) the key characteristics associated with nonresponse that are benchmarked and (b) the other characteristics imputed using CANCEIS.
   * The variability of estimates due to the adjustment procedure when different simulated censuses are taken (with a fixed true population). In the longer term this could be added with the variance estimation of whole estimation procedures to create measures of uncertainty.
4. Consideration of fall back options
   * The fall-back option if the CO method does not produce viable outputs will be to use the 2011 method, although we assess this risk to be low based on the work to date. However, should this event occur we will need to ensure the 2011 method is compatible with the new coverage estimation strategy and also to resolve the processing difficulties experience in 2011.
   * If the online census results in a higher proportion of missed people in counted households rather than from missed households then the two-stage approach may not be adequate. As contingency we would need to re-instate the imputation of persons from counted households in some form.

**References:**

Brown, J., Sexton, C., Taylor, A. and Abbott, O. (2011). Coverage Adjustment methodology for the 2011 Census.

<https://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/coverage-adjustment-methodology-for-the-2011-census.pdf>

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>

ONS (2011), “Item Edit and Imputation Evaluation”, 2011 Census Review and Evaluation Report.

<http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/evaluation---item-edit-and-imputation.pdf>

ONS (2012), “2011 Census Item Edit and Imputation Process”, 2011 Census Methods and Quality Report.

<http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/index.html>

ONS (2013), “2011 Census Assessment and Adjustment Evaluation”, 2011 Census Evaluation Report.

<http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/evaluation---coverage-assessment-and-adjustment.pdf>

Racinskij, V., (2018), “Coverage Estimation Strategy for the 2021 Census of England and Wales”, Internal paper.

Ross, H. and Abbott, O. (2015). “2021 Census Coverage Assessment and Adjustment strategy.”

Paper presented at the 30th GSS Methodology Advisory Committee, London, October 2015. Unpublished. Available on request.

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 (Statistics in Society) Volume 165, Issue 3, p 495–522.

Voas D. and Williamson P., (2000), “An Evaluation of the Combinatorial Optimisation

Approach to the Creation of Synthetic Microdata”. *International Journal of Population Geography*, Vol. 6, pp. 349-366.

**Appendix 1**

**Variables used in the modelling, calibration, and skeleton records**

**Table 1: Model variables**

|  |  |
| --- | --- |
| **Person** | **Household** |
| Age-sex group | Tenure |
| Hard to count index | Household structure |
| Tenure | Household ethnicity (England and Wales) |
| Household structure | Household religion (Northern Ireland) |
| Local authority | Hard to count index |
| Activity last week (adults only) | Local authority |
| Marital status (adults only) | Household born UK (England and Wales) |
| Ethnicity (England and Wales) | Household intention to stay |
| Religion (Northern Ireland) | Household address year ago |
| Address year ago |  |
| Born UK (England and Wales) |  |
| Born NI (Northern Ireland) |  |

**Table 2: Calibration variables**

|  |  |
| --- | --- |
| **Person-level weights** | **Household-level weights** |
| Age and sex | Household size |
| Hard to count index | Individual activity within household |
| Activity last week | Hard to count index |
| Tenure | Ethnicity within household |
| Ethnicity | Age-sex group within household |
|  | Household tenure |

**Appendix 2 - Use of the Combinatorial Optimisation (CO) for the census adjustment process**

**Tables 1a – 1e showing the total absolute error in estimates obtained from 1) running the CO method for the whole census adjustment process, 2) running the** **CO method for the missing households only (i.e. after missing persons have been imputed into counted households), and 3) the 2011 system.**

**Table 1a: East Midlands Estimation Area**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (number of categories) | TAE | | |
| CO for wholly missing households (average of 100 runs) | **CO for whole adjustment process (average over 100 runs)** | 2011 system |
| Age-Sex Group by LA (245) | 67 | **118** | 652 |
| Ethnicity (3) | 1 | **1** | 5132 |
| Activity Last Week (2) | 1 | **1** | 1130 |
| Household size (3) | 277 | **1** | 1106 |
| Hard to Count Index (2) | 0 | **0** | 396 |

**Table 1b: Inner London Estimation Area**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (number of categories) | TAE | | |
| CO for wholly missing households (average of 100 runs) | **CO for whole adjustment process (average over 100 runs)** | 2011 system |
| Age-Sex Group by LA (35) | 35 | **5** | 144 |
| Ethnicity (8) | 0 | **0** | 722 |
| Activity Last Week (2) | 0 | **0** | 112 |
| Household size (3) | 423 | **0** | 322 |
| Hard to Count Index (2) | 0 | **0** | 0 |

**Table 1c: Outer London Estimation Area**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (number of categories) | TAE | | |
| CO for wholly missing households (average of 100 runs) | **CO for whole adjustment process (average over 100 runs)** | 2011 system |
| Age-Sex Group by LA (35) | 38 | **69** | 103 |
| Ethnicity (10) | 0 | **0** | 427 |
| Activity Last Week (2) | 0 | **0** | 1169 |
| Household size (2) | 15 | **0** | 522 |
| Hard to Count Index (1) | 0 | **0** | 0 |

**Table 1d: South West Estimation Area**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (number of categories) | TAE | | |
| CO for wholly missing households (average of 100 runs) | **CO for whole adjustment process (average over 100 runs)** | 2011 system |
| Age-Sex Group by LA (245) | 82 | **89** | 705 |
| Ethnicity (2) | 0 | **0** | 141 |
| Activity Last Week (2) | 0 | **0** | 1599 |
| Household size (2) | 0 | **0** | 2148 |
| Hard to Count index (2) | 0 | **0** | 378 |

**Tables 1e: Yorkshire and the Humber Estimation Area**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (number of categories) | TAE | | |
| CO for wholly missing households (average of 100 runs) | **CO for whole adjustment process (average over 100 runs)** | 2011 system |
| Age-Sex Group by LA (70) | 25 | **6** | 310 |
| Ethnicity (5) | 0 | **0** | 112 |
| Activity Last Week (2) | 0 | **0** | 860 |
| Household size (3) | 0 | **0** | 475 |
| Hard to Count Index (2) | 0 | **0** | 5 |

Notes/Comments:

The TAE = Total Absolute Error which is the sum of the absolute differences between the observed counts and the expected (benchmark) counts for a given variable across each category.

In Oguz, S and Abbott, O., (2016) the focus is on running CO for the wholly missing households only and comparison of the results to the 2011 system.

**Table 2 Table showing total census counts before during and after coverage adjustment**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Estimation area | Unadjusted census database number of households | Unadjusted census database total people | Census database after people missed in counted households imputed, total people | Household shortfall | Unadjusted census database shortfall for number of one person households | Census database after people missed in counted households imputed, shortfall for number of one person households |
| East Midlands | 274,521 | 624,701 | 633,305 | 9441 | 1946 | 5056 |
| Inner London | 94,600 | 228,172 | 236,619 | 6608 | 1730 | 4171 |
| Outer London | 89,925 | 232,165 | 242,414 | 6994 | 1949 | 4446 |
| South West | 309,254 | 700,181 | 705,746 | 9607 | 2605 | 4420 |
| Yorkshire and the Humber | 242,987 | 550,208 | 562,343 | 12858 | 4434 | 9727 |

Note: Communal establishments are not included in the above table, these were imputed separately at a later stage.