The Proposed Duplication Calibration Method for the 2021 Census of England and Wales

Ceejay Hammond and Manuela Naprta, Office for National Statistics

Acknowledgements: Viktor Račinskij and James Brown

Commented by: Gareth Powell, Owen Abbott and Paul Smith

1. Introduction

The aim of the census is to enumerate every element within the target population. The census is subject to both missingness (undercoverage error) and incorrect inclusion (overcoverage error). For the 2021 Census, statistical methods have been developed to estimate population totals with the inclusion of undercoverage error and overcoverage error. Overcoverage occurs when a member of the population is enumerated more than once at the same location (duplicate census person response at the same location), more than once at a different location (duplicate person response at a different location), counted in the wrong location or is an erroneous person response (Račinskij and Hammond, 2019).

Both undercoverage and overcoverage are estimated using a 1-to-1 linkage exercise of the Census and the Census Coverage Survey (CCS). The aim of duplication calibration is to use an additional census to census linkage exercise to estimate the number of duplicates within the census. These estimates can then be used to calibrate the level of overcoverage in the census and with the proposed estimation methods, estimate population totals with higher precision.

2. 2011 Overcoverage estimation method

Duplicate returns from the same location and erroneous records, which can be identified, are resolved during data processing. Other types of overcoverage, however, remain in the census population and need to be accounted for during estimation.

The aim of the 2011 overcoverage estimation methodology was to measure the propensity for any individual, with a specific set of characteristics, to be incorrectly included in the census (ONS, 2012). These overcount propensities were then used to weight the dual system estimators, accounting for overcoverage present in the census. To achieve an estimate of overcount propensity, three major matching exercises were carried out:

- Localised census-to-CCS linkage – linking census and CCS records for the same people in the same postcode or a contiguous postcode (i.e. small area of geography)
- Regional census-to-CCS linkage – looking for matches that were not in the same small area of geography (i.e. individuals counted in the 'wrong' location).
- Regional census-to-census linkage – searching for duplicate census records, without the ability of identifying which one was 'correct' and which one was the duplicate record in the wrong location.

Overcount propensity

The overcount propensity, $\frac{1}{\gamma}$, where γ was defined by Large et al. (2011) as

$$\gamma_{\mathrm{ag}} = \frac{\text{Census count for group a in area g}}{\text{number of unique individuals for group a in area g correctly counted by the Census}}$$

and operationally denoted as

$$\gamma_{\mathrm{ag}} = \frac{\mathrm{X_{ag}}}{\mathrm{Y_{ag}}} = \frac{\mathrm{Y_{ag}} + \mathrm{E_{ag}}}{\mathrm{Y_{ag}}} \tag{1}$$

where the following three quantities are defined:

- $\mathrm{X_{ag}}$ is the total Census count for group a in area g,
- $\mathrm{Y_{ag}}$ is the correct Census count for group a in area g,
- $\mathrm{E_{ag}}$ is the erroneous (overcount) Census count for group a in area g.

Overcount groups, a, are 5 age-sex population groups. The areas, g, are Regions. These quantities are not observed directly and are, therefore, replaced with estimates to give

$$\hat{\gamma}_{\mathrm{ag}} = \frac{\hat{\mathrm{Y}}_{\mathrm{ag}} + \hat{\mathrm{E}}_{\mathrm{ag}}}{\hat{\mathrm{Y}}_{\mathrm{ag}}} \tag{2}$$

$\hat{\mathrm{Y}}_{\mathrm{ag}}$ is the design-weighted sum of the correct census returns found by the CCS in area g, while $\hat{\mathrm{E}}_{\mathrm{ag}}$ is the design-weighted sum, across the whole country, of incorrect returns in area g identified by the CCS sampling the individual at a different location.

An adjustment to (2) was made, recognising that $\hat{\mathrm{E}}_{\mathrm{ag}}$ includes overcount generated from both duplicate census returns from different locations, one of which is in the CCS not in area g, and returns from persons enumerated in the wrong location, by calibrating the CCS-based estimate of this quantity using duplicates found in the census-to-census matching exercise.

Duplication Calibration 2011 Census

For calibration, the quantity $\mathrm{D_{ahg}}$ was defined as the number of erroneous (overcount) Census returns for group a in area h within Region g, which are duplicates of the correct returns within the same Region. An estimate of $\mathrm{D_{ahg}}$ is made by using census-to-ccs linkage where the ccs location is used to identify the correct and incorrect half of the duplicate. For example, if CCS samples a postcode r within area j of Region g, and identifies the erroneous half of a duplicate within postcode q within area h of the same Region g, this will contribute to the estimate of $\mathrm{D_{ahg}}$. Therefore, $\mathrm{D_{ahg}}$ (3) is the weighted sum of erroneous duplicates in area h in Region g identified by the CCS sampling the *correct* half of the duplicate elsewhere in Region g.

$$\hat{\mathrm{D}}_{\mathrm{ahg}} = \sum_{\mathrm{q \in h}} \sum_{\mathrm{j \in g}} \sum_{\mathrm{r \in CCS_j}} \mathrm{w_{rjg} D_{aqhg,arjg}} \tag{3}$$

where the following three quantities are defined:

- $\mathrm{w_{rjg}}$ is the CCS sampling weight for postcode r in area j within Region g
- $\mathrm{D_{aqhg,arjg}}$ is a census return in a CCS sample, in postcode r within area j of Region g and has a duplicate (incorrect half) within postcode q in area h in Region g
- $\hat{\mathrm{D}}_{\mathrm{ag}} = \sum_{\mathrm{h \in g}} \hat{\mathrm{D}}_{\mathrm{ahg}}$ is an estimate of the total number of erroneous duplicates across Region g

In the case of triplications, we expect there will be very few cases of this so do not need to be resolved here and should be resolved during Census data processing.

In principle, this should equal the estimated number of matched pairs found using inverse sampling and census-to-census matching, $\hat{\mathrm{P}}_{\mathrm{ag}}$. The estimate of $\mathrm{P_{ag}}$ was derived using the inverse sampling method (Haldane, 1945) applied to population strata targeted for overcount (e.g. students). A minimum sample of 5000 records is drawn for each stratum, within each Region (10 regions in England and Wales). Duplicates are searched for until a suitable number is found, with additional samples being drawn if that threshold value of duplicates is not reached initially. A detailed description of the methods can be found in Abbott and Large (2009).

However, as $\hat{\mathrm{D}}_{\mathrm{ag}} \neq \hat{\mathrm{P}}_{\mathrm{ag}}$, we make the assumption that the census-to-census matching gives a more precise estimate of the number of duplicates within Region g. This is due to the high linkage precision from census-to-census linkage, targeted linkage within each group in each region and the requirement for the coefficient of variance CV(p) to be less than 10%.

With this we can improve the CCS-based estimate of overcount by using a ratio estimator, to give

$$\tilde{\mathrm{D}}_{\mathrm{ahg}} = \sum_{\mathrm{q \in h}} \sum_{\mathrm{j \in g}} \sum_{\mathrm{r \in CCS_j}} (\mathrm{w_{rjg}} * {}^{\hat{\mathrm{P}}_{\mathrm{ag}}}\!/_{\hat{\mathrm{D}}_{\mathrm{ag}}}) * \mathrm{D_{aqhg,arjg}} \tag{4}$$

This adjusted weight $(\mathrm{w_{rjg}} * {}^{\hat{\mathrm{P}}_{\mathrm{ag}}}\!/_{\hat{\mathrm{D}}_{\mathrm{ag}}})$can then be used in the estimation of $\mathrm{E_{ag}}$ to be used in the (2). A key assumption is that the ratio [calibration] adjustment for duplication overcount is a good predictor of the adjustment for wrong location overcount.

Adjusting the DSE for overcount

The overcount propensities, $\hat{\gamma}_{\mathrm{ag}}$, calculated following the methods outlined above were then incorporated into the dual system estimators at the level of a cluster of postcodes, s, within groups, a. The dual-system estimator (DSE) of a true population, $\mathrm{N_{as}}$, is defined as

$$\hat{\mathrm{N}}_{\mathrm{as}} = \frac{\mathrm{Z_{as} \times Y_{as}}}{\mathrm{M_{as}}} \tag{5}$$

where the following three quantities are defined:

- $\mathrm{Z_{as}}$ is the CCS count for group a, in cluster of postcodes s,
- $\mathrm{Y_{as}}$ is the census count of correct individuals for group a, in cluster of postcodes s,
- $\mathrm{M_{as}}$ is the matched count for group a, in cluster of postcodes s

In practice, the quantity $\mathrm{X_{as}}$, which is the census count containing overcount, is observed, giving the DSE as:

$$\hat{\mathrm{N}}_{\mathrm{as}}^{\mathrm{C}} = \frac{(\mathrm{Z_{as}}+1) \times \left({}^{\mathrm{X_{as}}}\!/_{\hat{\gamma}_{\mathrm{ag}}} + 1 \right)}{(\mathrm{M_{as}}+1)} - 1 \qquad (6)$$

(6) includes the Chapman Correction (superscript C) as the DSE was applied to small populations. Acknowledging that the census count cannot be adjusted directly but that, assuming the underlying overcount propensity, $\hat{\gamma}_{\mathrm{ag}}$, is constant across area g then, $\mathrm{E}\left[{}^{\mathrm{X_{as}}}\!/_{\hat{\gamma}_{\mathrm{ag}}} \mid \mathrm{X_{as}} \right] \cong \mathrm{Y_{as}}$ and therefore, $\tilde{\mathrm{N}}_{\mathrm{as}}$ is approximately unbiased for $\hat{\mathrm{N}}_{\mathrm{as}}$.

A more comprehensive description of the DSE adjustment for estimated overcount and overcount propensity calculations can be found in Large et al., (2011).

3. 2021 Under- and Overcoverage estimation strategy

Undercoverage overview

As in the 2011 Census, a CCS will be undertaken, and the CCS data will be linked to the census data for under- and overcoverage estimation. It is expected that census data for the entire country will be available for estimation sooner than in 2011, due to the smaller number of paper returns. Batching Local Authorities into Estimation Areas will not be necessary, and estimation can be carried out at a national or regional scale, similarly to the approach used in the 2010 US Census (US Census Bureau, 2008, 2012). Census coverage for a relatively large set of characteristics will be estimated using a modelling approach, with logistic and mixed effects logistic regression models (Alho, 1990; Alho et al., 1993; Chamber & Clark, 2012), contingent on a sufficiently large sample of data.

The aim is to estimate the population size for a domain of interest, consisting of individuals or households with certain characteristics, along with some geographical property. For example, an age-sex group a, in an area l (e.g. a local authority). For a population element, i, the estimated census response probability is either $\hat{\pi}_{\mathrm{i}} = \frac{1}{1+\mathrm{e}^{\left(-[\hat{\beta}_0+\sum_{\mathrm{k}=1}^{\mathrm{K}}\hat{\beta}_{\mathrm{k}}\mathrm{x_k}]\right)}}$, for a logistic regression model, or $\hat{\pi}_{\mathrm{i}} = \frac{1}{1+\mathrm{e}^{\left(-[\hat{\beta}_0+\sum_{\mathrm{k}=1}^{\mathrm{K}}\hat{\beta}_{\mathrm{k}}\mathrm{x_k}+\hat{u}_l]\right)}}$, for a mixed effects logistic regression model. Estimated census non-response weights (reciprocals of estimated census response probabilities) for an individual or household with certain characteristics can be applied to each census observation with matching characteristics (US Census Bureau, 2012). Weighted census observations can then be summed to produce an estimated population size of elements with the characteristic (8) (9); see Račinskij (2018) for details.

$$\hat{\mathrm{T}}_{\mathrm{al}}^{\mathrm{LR}} = \sum_{\mathrm{i}\in\mathrm{al}} \left[\frac{1}{1+\ \mathrm{e}^{\left(-[\hat{\beta}_0+\sum_{\mathrm{k}=1}^{K}\hat{\beta}_k \mathrm{x}_k]\right)}} \right]^{-1} \qquad (8)$$

$$\widehat{\mathrm{T}}_{\mathrm{al}}^{\mathrm{MELR}} = \sum_{\mathrm{i}\in\mathrm{al}} \left[\frac{1}{1 + \ \mathrm{e}^{\left(-[\widehat{\upbeta}_0 + \sum_{\mathrm{k}=1}^{K} \widehat{\upbeta}_k \mathrm{x}_k + \widehat{\mathrm{u}}_\mathrm{l}]\right)}}\right]^{-1} \tag{9}$$

A vector of k covariates, x, is used and optionally a random effect of area l (e.g. a local authority) in the models for estimation.

Overcoverage overview

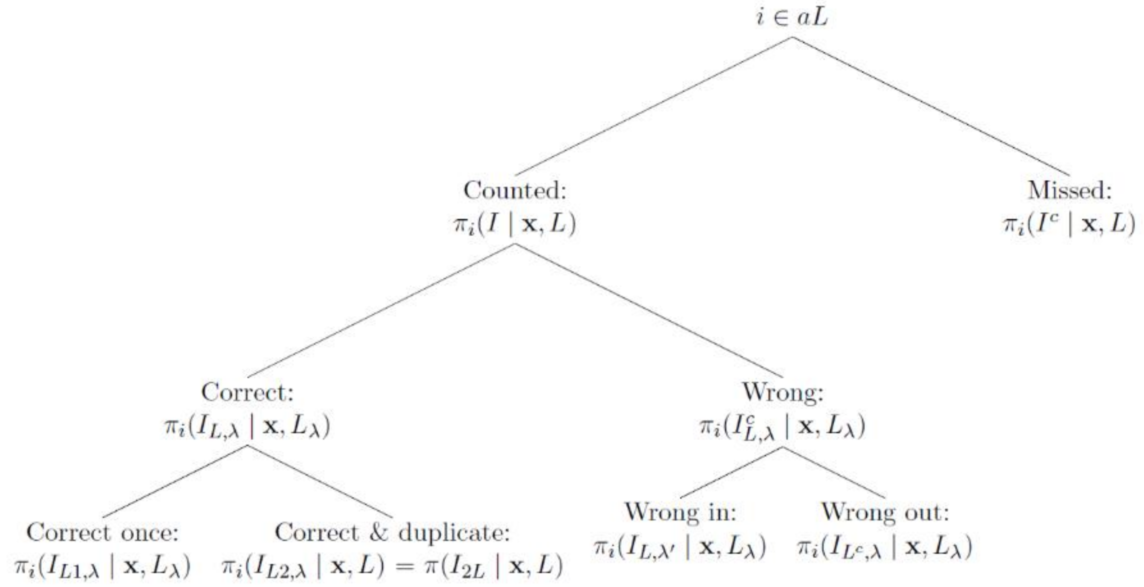
The method proposed for the 2021 census overcoverage estimation reflects that used in the undercoverage estimation. Every member of the population has a correct location associated with them, however, in practice it is not always necessary to determine this location to an atomic level. Therefore, in the 2021 overcoverage estimation strategy (Račinskij and Hammond, 2019), a search area, λ, was defined. The search area is a small region 'around' the atomic correct location (captured by the CCS) such that a person with a census return within that search area is classified as correctly enumerated even if the exact location is incorrect. This search area, λ, belongs to a larger area L, e.g. a local authority, and we denote a search carried out in L, at the level λ, as L_λ.

As in 2011, data processing will handle overcount from duplicates within the same household, recognisable erroneous returns as well as returns from members of a non-target population. Overcount returns from individuals enumerated in the wrong location (i.e. outside λ), whether duplicates or records solely from the wrong location, will remain in the census population.

To detect overcount cases, the Census-to-CCS linkage process is used, which satisfies a 1-to-1 constraint. Therefore in the case of duplicates, a CCS record will link to two Census records, where the status of the duplicate Census record will be determined by whether is location is the same as the corresponding CCS record (correct part of duplicate) or different to the corresponding CCS record (incorrect part of duplicate). For wrong location overcount cases, these will be determined if their location does not match their linked CCS records location.

Recall the response generation for the simulation study (Račinskij and Hammond, 2019), which outlines the response of elements (i) within the target population.

Figure 1: Census Response Generation for the Simulation Study


$i \in aL$
Counted:
$\pi_i(I \mid \mathbf{x}, L)$
Missed:
$\pi_i(I^c \mid \mathbf{x}, L)$
Correct:
$\pi_i(I_{L,\lambda} \mid \mathbf{x}, L_\lambda)$
Wrong:
$\pi_i(I^c_{L,\lambda} \mid \mathbf{x}, L_\lambda)$
Correct once:
$\pi_i(I_{L1,\lambda} \mid \mathbf{x}, L_\lambda)$
Correct & duplicate:
$\pi_i(I_{L2,\lambda} \mid \mathbf{x}, L) = \pi(I_{2L} \mid \mathbf{x}, L)$
Wrong in:
$\pi_i(I_{L,\lambda'} \mid \mathbf{x}, L_\lambda)$
Wrong out:
$\pi_i(I_{L^c,\lambda} \mid \mathbf{x}, L_\lambda)$

True population values are denoted as covariates x and location L. A superscript r is added for those reported in the census, where x^r are the characteristics that the record reported in the census and L^r is the location of the record reported in the census. It is assumed that in the census there are no errors reporting x, whereas there are errors reporting L. Any error in the reporting of L^r_λ contributes to overcount. If there was no overcount (an assumption of the DSE), then $L^r_\lambda = L_\lambda$, i.e. the location reported in the census and the true location are the same. The estimator for the population total (8) can, in this case, be given as

$$\hat{T}_{aL,reg} = \sum_{i \in aL^r} \frac{1}{\hat{\pi}_i(m \mid x, L^r_\lambda)} \tag{10}$$

Where $\hat{\pi}_i(m \mid x, L^r_\lambda)$ is the estimated probability that a randomly selected CCS record in L_λ matches (m) a census record in that area. In the presence of overcount this expression the estimator (10) is multiplied by the estimated probability of an element being correctly enumerated.

$$\hat{T}^{(oc)}_{aL,reg} = \sum_{i \in aL^r} \frac{\hat{\pi}_i(ce \mid x^{(oc)}, L^r_\lambda)}{\hat{\pi}_i(m \mid x, L^r_\lambda)} \tag{11}$$

where $\hat{\pi}_i(ce \mid x^{(oc)}, L^r_\lambda)$ is the estimated probability that an element is correctly enumerated in the Census. That is, the probability that an observed return in L_λ, with characteristics $x^{(oc)}$, is a return of an element that truly resides in L_λ.

The 2021 overcount probabilities are analogous to the 2011 overcount propensities as they will be used in a similar way to down-weight the census counts. However, each census element will now have an individual probability (given their characteristics), instead of being applied a propensity calculated for a larger group they belong to (i.e. age-sex and region groups).

4. Duplication Calibration 2021 Census

2021 Census to Census Linkage

Introduction

Census to census linkage is the process of detecting people that have been enumerated in the census more than once. As in 2011, a probabilistic matching algorithm (Shipsey and White, 2020) will be used to generate candidate duplicate pairs alongside an automated checking algorithm and Inverse sampling.

Automated checking algorithm

This probabilistic matching algorithm generates many candidate pairs, all of which require clerical reviewing to ensure whether the duplicate is genuine or not. To reduce the amount of clerical review necessary in the 2021 Census, the linkage team has created an automated checking algorithm (Shipsey and White, 2020), attempting to replicate the reasons behind classifying a candidate pair as genuine or not, in distinct cases. This has allowed candidate pairs to be split into three categories:

- Accept automatically as a duplicate
- Reject automatically as a duplicate
- Send to clerical review

Inverse sampling method

To estimate the prevalence of overcount in different population groups the inverse sampling method (Haldane, 1945) of 2011 combined with the automated checking algorithm (Shipsey and White, 2020) will be used. The inverse sampling technique involves taking, initially, a random sample of 5000 census records, from each of the 16 population groups in each of the regions, until 102 duplicates are found in each group. This method ensures an estimation of the proportion of individuals counted more than once, with good relative error, as it is expected that P will be small (P < 0.01). The number 102 was selected to give a coefficient of variance CV(p) of less than 10%.

$$\hat{P} = \frac{m-1}{n-1} \tag{12}$$

Where m is the pre-determined threshold number of duplicates, and n is the sample size at which the number of duplicates m was reached.

The following 16 population groups from each of the 11 regions will be sampled for duplicates in 2021, in priority order:

- Persons who have indicated they have a second residence on the census
- Students aged 18 to 25 by gender (2 groups)
- Persons who have indicated their address one year ago is not the same as their current address
- Armed forces personnel
- Adults enumerated at a communal establishment aged 16-44, 45-74 and 75+ (3 groups)
- Children aged 0-4, 5-15 (2 groups)
- Individuals who completed the questionnaire on paper aged 16-29, 30-49 and 50+ (3 groups)
- Everyone else by broad age groups 16-29, 30-49, 50+ (3 groups)

These groups are hierarchical and mutually exclusive, for example, any individual that indicates a second address is considered part of the first group only, despite also fitting into subsequent groups. The groups and their order have been determined using evidence from the longitudinal study (ONS, 2014) of overcount in certain population groups, as well as the results from the 2011 Census to Census matching (ONS, 2012) and findings from a clerical review during the development of the automatic checking algorithm. More information about the selection of the groups in this list can be found in Shipsey and White (2020).

In summary, the census to census linkage process is as follows:

- Census population is split into 11 regions put into the 16 overcount groups given above to be sampled.
- For each group in every region a minimum random sample of 5000 records is taken.
- The probabilistic matching algorithm generates candidate pairs for each sample across the Census.
- Candidate pairs are sent through to the automatic checking algorithm.
- Once the number of automatically accepted candidate pairs plus clerically confirmed matches is 102 or above, the process stops.
- Otherwise, another sample is selected, adjusting the size according to the proportion of duplicates already found, and the process is repeated until 102 duplicates are found.

Proposed Duplication Calibration method

The proposed duplication calibration method developed by James Brown (2019) is outlined below. For the population of interest, N an element (in this case, a person) i is either correctly or incorrectly enumerated in the Census. Using the Census-to-CCS 1-to-1 linkage exercise we define,

- Y = 0 used to identify census records that were correctly enumerated
- Y = 1 to identify census records that were enumerated in the wrong location
- Y = 2 to identify census records that were duplicates of an element enumerated elsewhere in the census

The probabilities of these outcomes are as follows,

- Let π_{0i} be the probability of an element i of the population, N being correctly enumerated in the census
- Let π_{1i} be the probability of an element i of the population, N being enumerated in the wrong location in the census
- Let π_{2i} be the probability of an element i of the population, N being a duplicate of an element enumerated elsewhere in the census

Correct enumeration model where an element i is either correctly enumerated (Y=0) or incorrectly enumerated (Y=1 or Y=2) in the census:

$$\ln\left(\frac{\pi_{0i}}{1-\pi_{0i}}\right) = \left(\frac{\pi_{0i}}{\pi_{1i}+\pi_{2i}}\right) = x_i^T\beta_0 \tag{13}$$

where $\pi_{1i} + \pi_{2i}$ is the total probability of an element i in the census being incorrectly enumerated (overcount).

The estimated probability of being correctly enumerated is

$$\hat{\pi}_{0i} = \frac{e^{x_i^T\hat{\boldsymbol{\beta}}_0}}{1+e^{x_i^T\hat{\boldsymbol{\beta}}_0}} \qquad (14)$$

Now subset all identified (census-to-ccs) overcount cases to model the probability of being a wrong location return among the overcount population n (where n is a subset of N) , where they can be identified as either Y = 1 or Y = 2 in the data.

Wrong location model (dependent variable Y = 1), where an element i is either incorrectly enumerated in the wrong location (Y=1) or incorrectly enumerated as the wrong part of a duplicate (Y=2):

$$\ln\left(\frac{\pi_{1i}}{1-\pi_{1i}}\right) = \left(\frac{\pi_{1i}}{\pi_{2i}}\right) = x_i^T\beta_1, \qquad (15)$$

Therefore, for all elements i in population n, the estimated probability of being enumerated in the wrong location is

$$\hat{\pi}_{1i} = \frac{e^{x_i^T\hat{\boldsymbol{\beta}}_1}}{1+e^{x_i^T\hat{\boldsymbol{\beta}}_1}} \qquad (16)$$

The estimated odds ratio of (13),

$\left(\frac{\hat{\pi}_{1i}}{\hat{\pi}_{2i}}\right) = e^{x_i^T\beta_1}$ and therefore $\hat{\pi}_{1i} = \hat{\pi}_{2i}e^{x_i^T\beta_1}$ (17)

Recall the total estimated overcount probability is $\hat{\pi}_{1i} + \hat{\pi}_{2i} = \frac{1}{1+e^{x_i^T\hat{\boldsymbol{\beta}}_1}}$ (18)

By substituting (17) into (18). The estimated probability of being a duplicate is

$$\hat{\pi}_{2i} = \frac{1}{(1+e^{x_i^T\hat{\boldsymbol{\beta}}_1})(1+e^{x_i^T\hat{\boldsymbol{\beta}}_0})} \qquad (19)$$

To estimate the number of duplicates over the domain of interest (16 groups within each region) within the census from the census-to-ccs linkage exercise, $\widehat{D}_{ccs} = \sum_i \hat{\pi}_{2i}$

In addition to this, we also have an estimated number of duplicates over the domain of interest from the census-to-census linkage study, $\widehat{D}_{cen}$

Therefore, the calibrated estimated probability of an element being a duplicate is as follows,

$$\tilde{\pi}_{2i} = \frac{\widehat{D}_{cen}}{\widehat{D}_{ccs}}\hat{\pi}_{2i} \qquad (20)$$

This calibration method will then be used to produce calibrated estimates of the probability of an element i being correctly enumerated. Recall $\hat{\pi}_{0i} = 1 - (\hat{\pi}_{1i} + \hat{\pi}_{2i})$, therefore,

$$\tilde{\pi}_{0i} = 1 - \frac{\widehat{D}_{cen}}{\widehat{D}_{ccs}}(\hat{\pi}_{1i} + \hat{\pi}_{2i}) \qquad (21)$$

Therefore, the calibrated estimator for the population total (11) can, in this case, be given as

$$\widetilde{\mathrm{T}}^{(oc)}_{\mathrm{aL,reg}} = \sum\nolimits_{\mathrm{i}\in\mathrm{aL}^{\mathrm{r}}} \frac{\tilde{\pi}_{\mathrm{i}}(\mathrm{ce} \mid \mathrm{x}^{(oc)},\mathrm{L}^{\mathrm{r}}_{\lambda})}{\hat{\pi}_{\mathrm{i}}(\mathrm{m} \mid \mathrm{x},\mathrm{L}^{\mathrm{r}}_{\lambda})} \qquad (22)$$

In summary, the duplication calibration method is as follows:

- Model the probability of an element i being correctly enumerated in the Census where the population N, are those from the Census-to-CCS linkage.

- Model the probability of an element i being in the wrong location in the Census from the incorrect enumeration population, n (subset of population N).
- Estimate the probability of each element being a duplicate (19).
- Sum up these estimated duplicate probabilities across each of the 16 groups within each region (outlined above) $\widehat{D}_{ccs}$.
- From census-to-census linkage estimate the number of duplicates in each group within each region (outlined above) $\widehat{D}_{cen}$.
- Duplication calibration ratio $\frac{\widehat{D}_{cen}}{\widehat{D}_{ccs}}$.
- Apply the duplication calibration ratio to the probability of being incorrectly enumerated in the census.

This approach builds on the 2011 duplication calibration approach in terms of establishing the true level of duplication in the census, however, it is integrated within the proposed coverage estimation methods. This method requires additional model selection for the wrong location model (15).

5. Simulation Study

Introduction

A small simulation study is used to assess the performance of the proposed duplication calibration approach. This was done by estimating population totals at both national and age-sex by local authority levels when duplication calibration is not included (11) and is included (22) alongside the proposed estimation approach for the 2021 Census of England and Wales. In the study a population of approximately 3 million individuals (nation level) was used, from a combination of two small regions, where within each region eight local authorities were paired together to give a total of eight local authority across the two regions (four local authority within each region). There were four age-sex groups and two hard-to-count groups. Our domains of interest were both age-sex by local authority and national levels.

In this simulation study, three scenarios were assessed. Although this simulation is complex, some of the complexities we would expect to see in the 2021 Census are not reflected. Therefore, the simulation study aids the estimators in performing well. An example of this, is linkage error which is not introduced into this simulation study. To introduce bias into the simulation study, which allows for the duplication calibration method to correct, a simple correct enumeration (13) and wrong location (15) model were used (see appendix for results where bias was not introduced).

For each scenario outlined (Table 1) two hundred censuses were generated where the level of undercoverage was unchanged and the level of overcoverage varied across the scenarios, where scenario 1 had the largest level of overcoverage and the largest level of duplication, whereas scenario 3 had the lowest level of overcoverage and the lowest level of duplication. The response rate for the census coverage survey (*S*) remained constant.

Table 1: Simulation parameters for three scenarios (log-odds scale)

Scenario	Event	Intercept	Age-sex group 1	Age-sex group 2	Age-sex group 3	Age-sex group 4	HtC 1	HtC 2	Region 1	Region 2	Local Authority ID 1	Local Authority ID 2	Local Authority ID 3	Local Authority ID 4	Local Authority ID 5	Local Authority ID 6	Local Authority ID 7	Local Authority ID 8
s01	I	2.3	-0.3	-0.6	0.6	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s01	$I_{L,\lambda}$	4	-0.5	-0.6	0.1	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s01	I_{2L}	-4.8	0.5	0.6	-0.1	0	-0.5	0	-0.5	0	0	0	0	0	0	0	0	0
s01	$I_{L,\lambda'}$	0	-0.5	-0.6	0.1	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s01	S	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s02	I	2.3	-0.3	-0.6	0.6	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s02	$I_{L,\lambda}$	4	-0.5	-0.6	0.1	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s02	I_{2L}	-5.4	0.5	0.6	-0.1	0	-0.5	0	-0.5	0	0	0	0	0	0	0	0	0
s02	$I_{L,\lambda'}$	0	-0.5	-0.6	0.1	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s02	S	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s03	I	2.3	-0.3	-0.6	0.6	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s03	$I_{L,\lambda}$	5.1	-0.5	-0.6	0.1	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s03	I_{2L}	-6	0.5	0.6	-0.1	0	-0.5	0	-0.5	0	0	0	0	0	0	0	0	0
s03	$I_{L,\lambda'}$	1.3	-0.5	-0.6	0.1	0	0.5	0	0.5	0	0	0	0	0	0	0	0	0
s03	S	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Results

The results show the Relative Bias, Relative Root Mean Squared Error and Relative Standard Error for each of the outlined scenarios (Table 1) for the domain of interest, age-sex by local authority level (Figures 2 to 10) and the Total Relative Bias, Relative Root Mean Squared Error and Relative Standard Error for each scenario at national level (Tables 2 to 4).
For each scenario the main effects used in the undercoverage model (10) were Region, HtC and Age-sex group, the correct enumeration model (13) were Region and Student and the wrong location model (15) were Region and Student (for scenarios 1 and 2) and Region (for scenario 3).

Figure 2: Relative Bias for Age-sex by Local Authority for Simulation Scenario 1, ordered by Age-sex by Local Authority

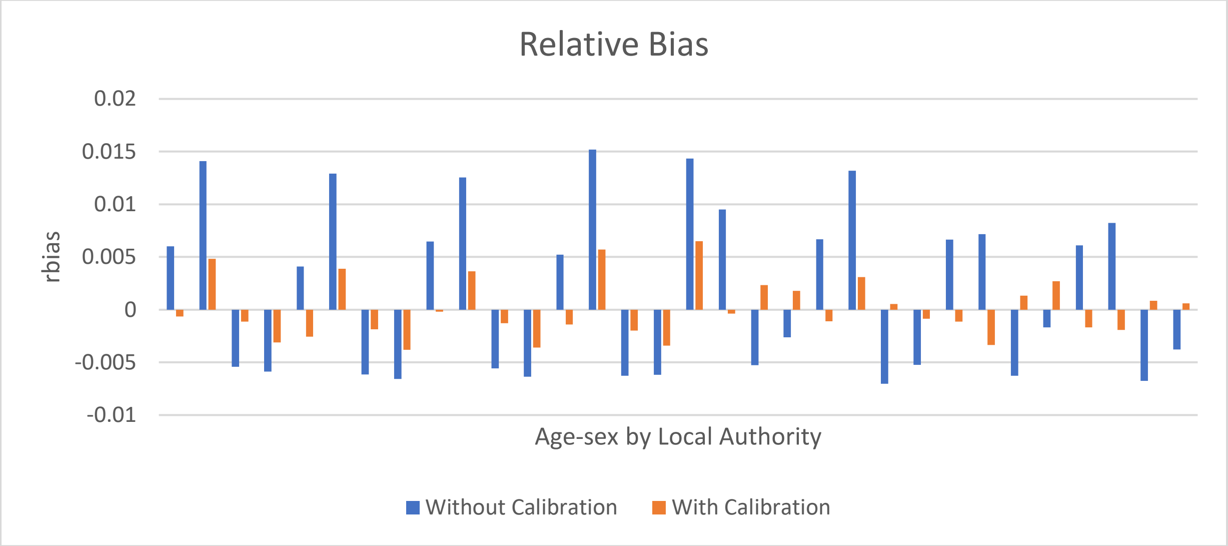


Figure 3: Relative Bias for Age-sex by Local Authority for Simulation Scenario 2 ordered by Age-sex by Local Authority

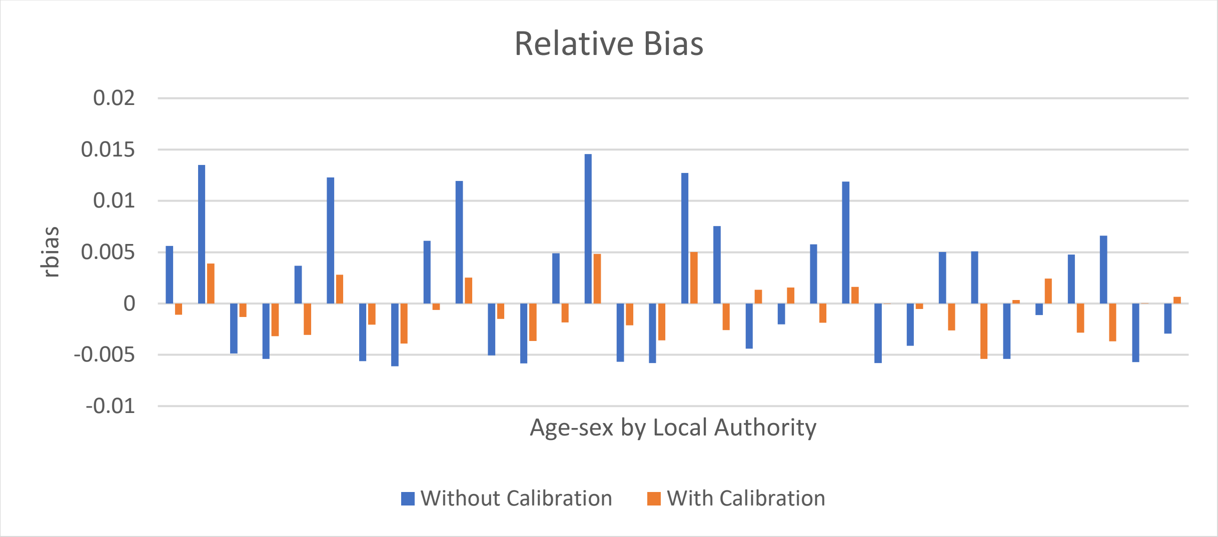
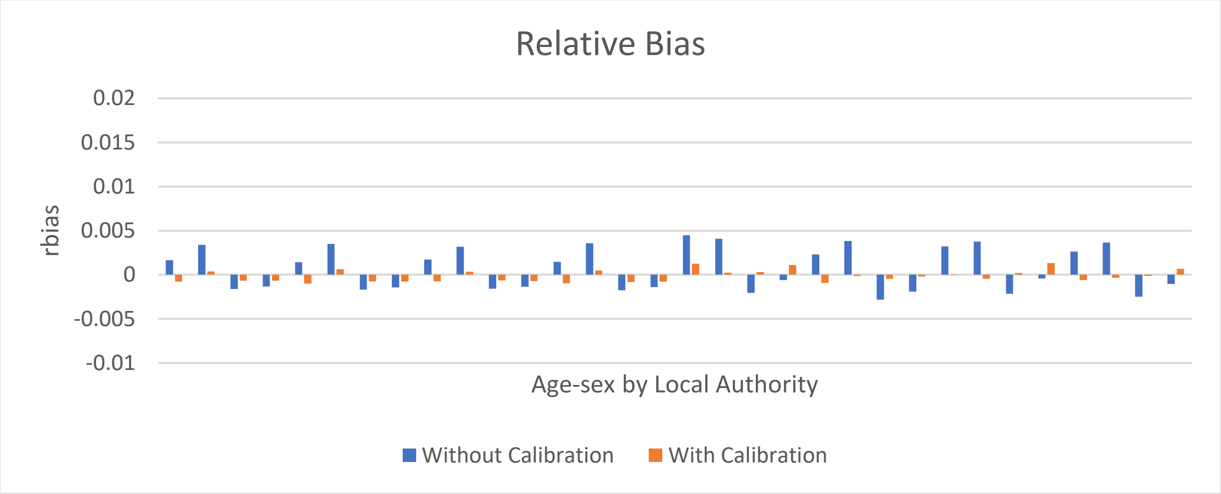


Figure 4: Relative Bias for Age-sex by Local Authority for Simulation Scenario 3 ordered by Age-sex by Local Authority



For each scenario, the relative bias for all domains of interest where the estimator includes duplication calibration (22) is smaller (in absolute terms) than for when the estimator does not include duplication calibration (11). This implies that the estimator that includes duplication calibration outperforms the estimator where it is not included.

Figure 5: Relative Root Mean Squared Error for Age-sex by Local Authority for Simulation Scenario 1 ordered by Age-sex by Local Authority

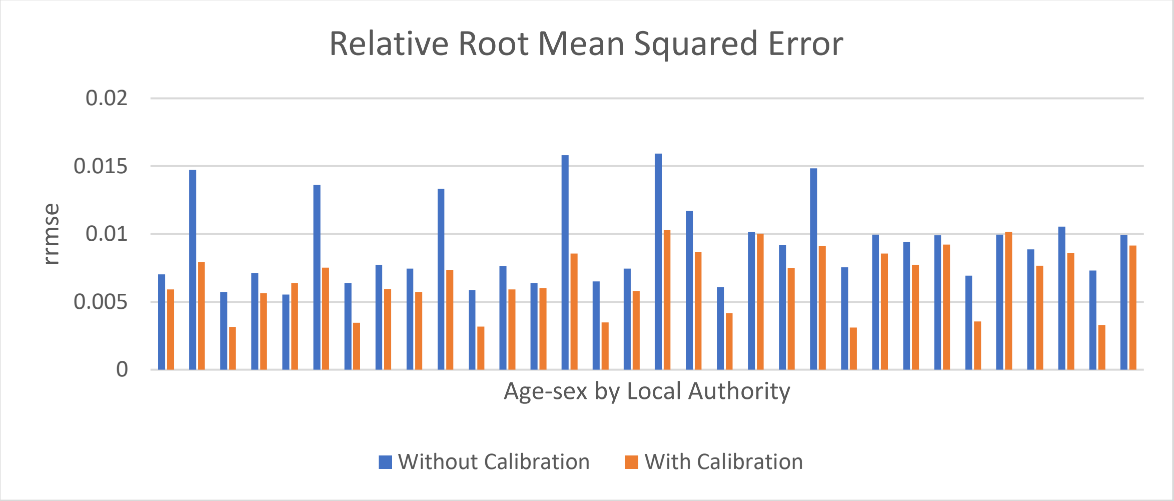


Figure 6: Relative Root Mean Squared Error for Age-sex by Local Authority for Simulation Scenario 2 ordered by Age-sex by Local Authority

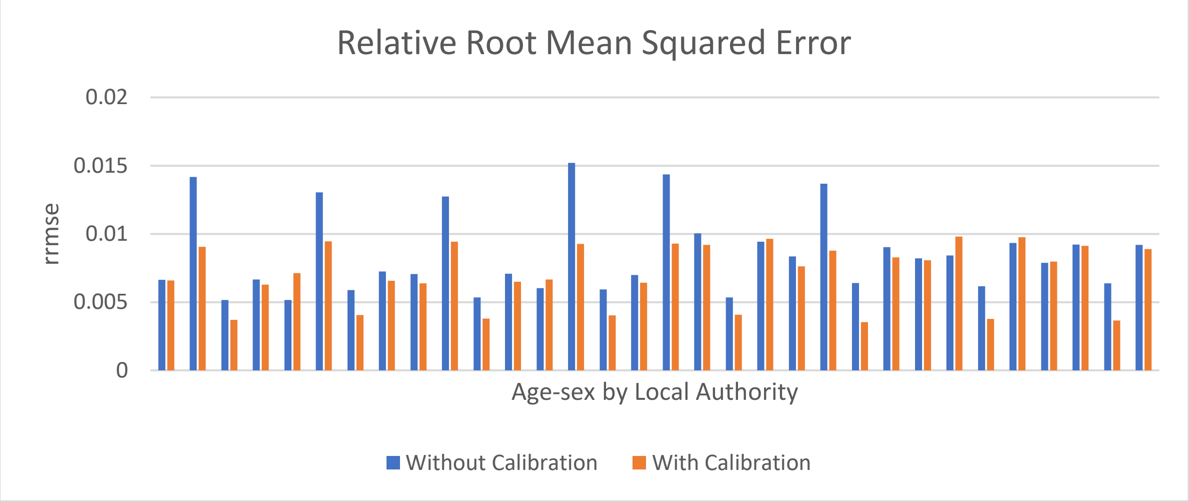
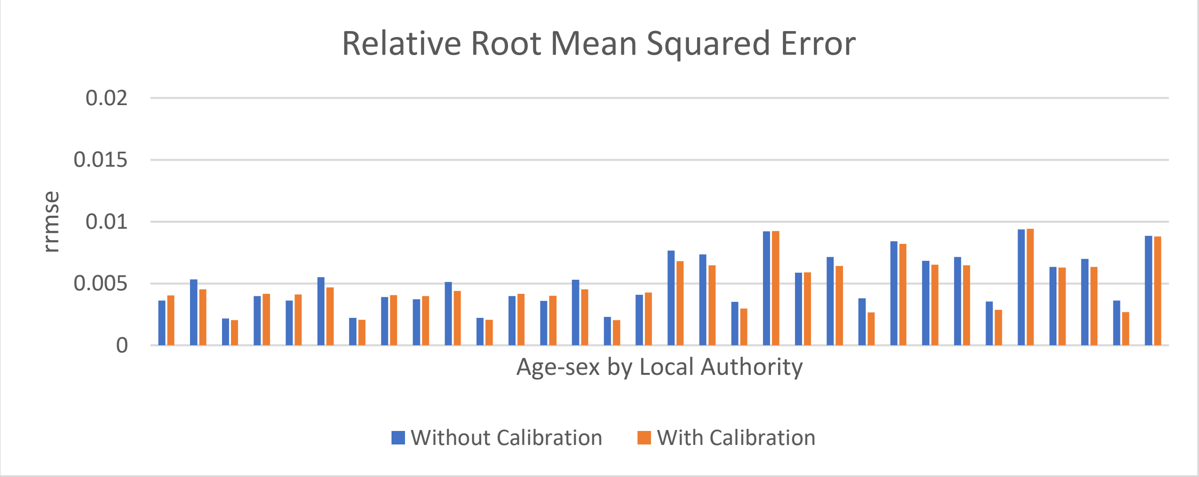


Figure 7: Relative Root Mean Squared Error for Age-sex by Local Authority for Simulation Scenario 3 ordered by Age-sex by Local Authority



For each scenario, the relative root mean squared error is generally smaller across the domains of interest when the estimator (22) is used. Overall, the estimator that includes duplication calibration (22) outperforms the estimator where it is not included (11). These results show the estimator (22) is robust at calibrating duplicates as for scenario 3 where the level of overcount and duplication is the smallest across all of the scenarios, the rrmse values do not differ between estimators (11) and (22) across age-sex by local authorities.

Although the estimator with duplication calibration (22) outperforms the estimator where there is no duplication calibration (11) it is clear to see that both estimators perform well at estimating population totals where the highest value of relative bias across the 3 scenarios is 1.5%.

Figure 8: Relative Standard Error for Age-sex by Local Authority for Simulation Scenario 1 ordered by Age-sex by Local Authority

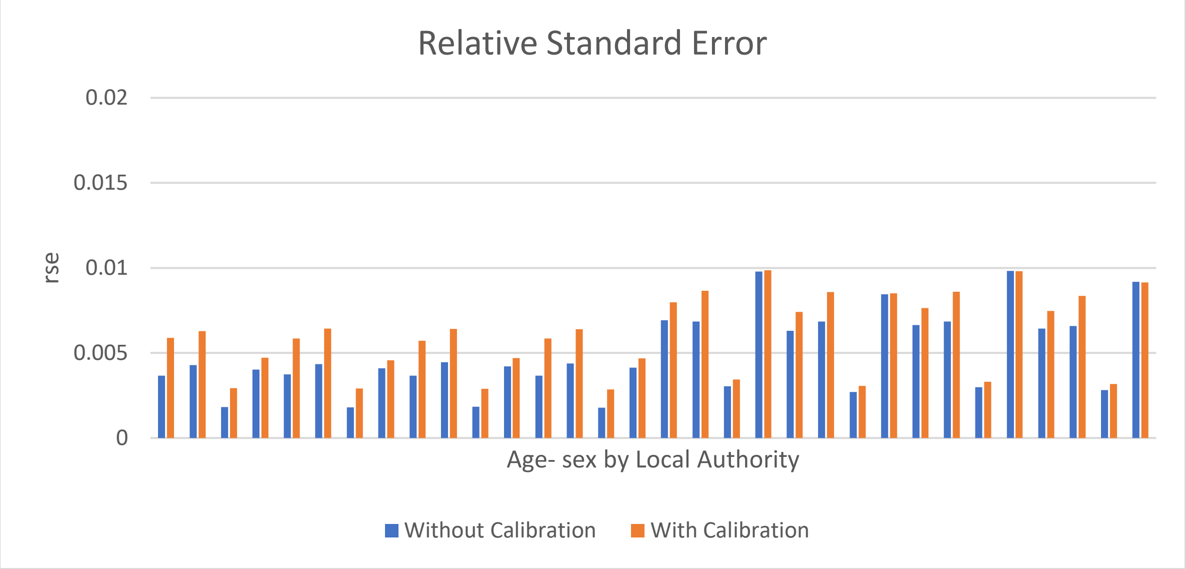


Figure 9: Relative Standard Error for Age-sex by Local Authority for Simulation Scenario 2 ordered by Age-sex by Local Authority

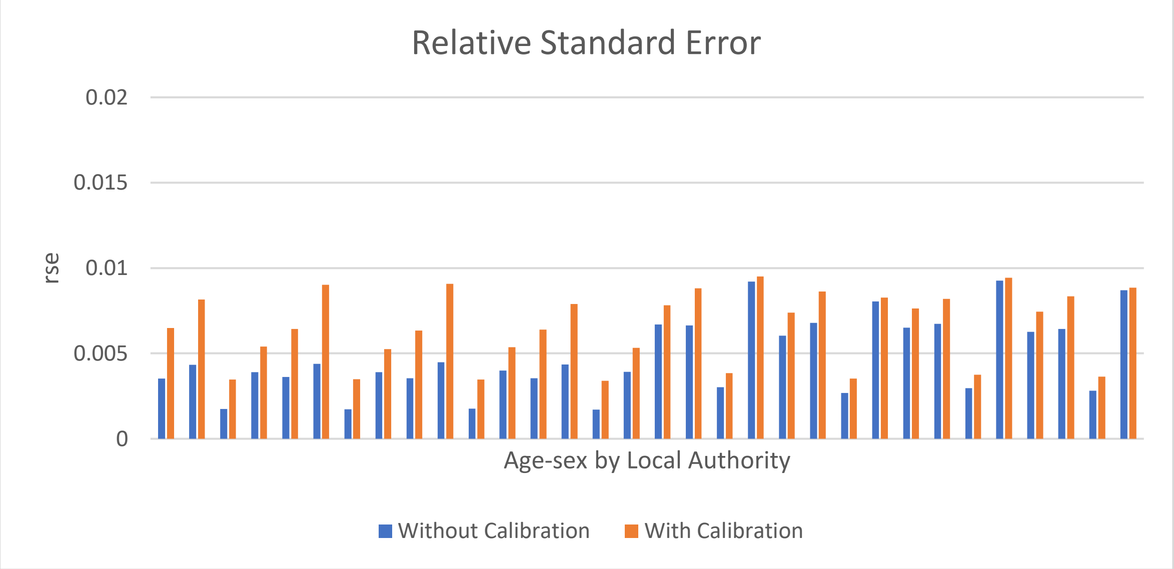
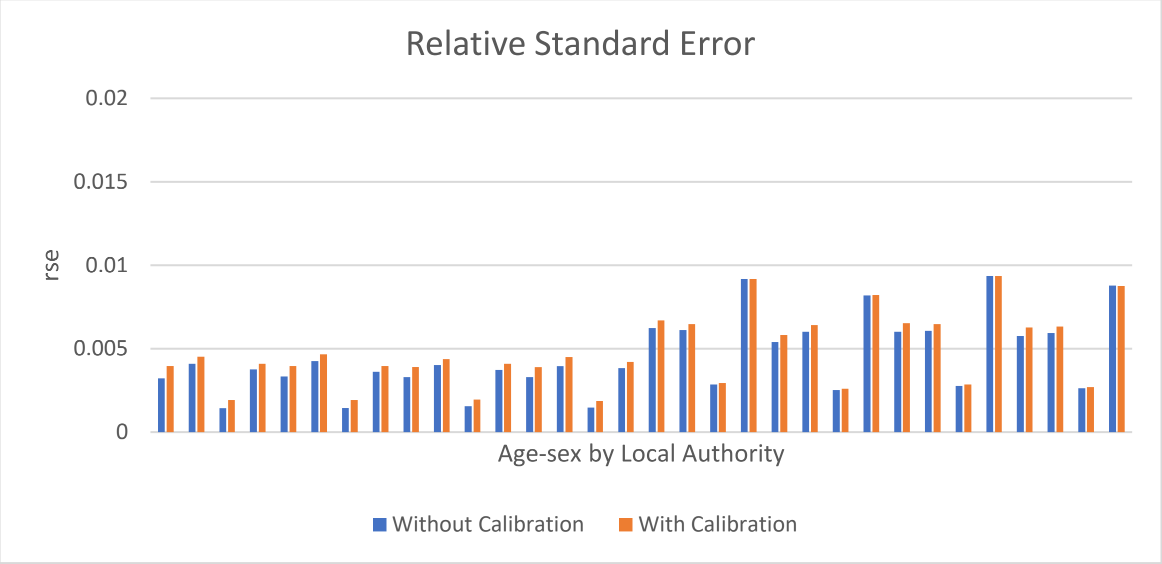


Figure 10: Relative Standard Error for Age-sex by Local Authority for Simulation Scenario 3 ordered by Age-sex by Local Authority



For each scenario across all the domains of interest the relative standard error is larger when the estimator that includes duplication calibration (22) is used, than when it is not (11). This suggests that there is a larger variance of the population estimates across the 200 censuses, as expected from the addition of the wrong location model and duplication calibration to the estimator, which is derived from the census-to-census linkage process, where the coefficient of variation of the estimated number of duplicates is less than 10%, which introduces variance. The proposed 2021 estimators have relatively low variance compared to the DSE estimator used in 2011.

Table 2: Total Relative Bias at national level for each Simulation Scenario

Scenario	Without Calibration	With Calibration
s01	0.000431254	-0.000017249
s02	0.00035668	-0.000847604
s03	0.000227782	-0.000255721

Table 3: Total Relative Root Mean Squared Error at national level for each Simulation Scenario

Scenario	Without Calibration	With Calibration
s01	0.001801147	0.002404252
s02	0.00175476	0.002887212
s03	0.001701479	0.001919126

Table 4: Total Relative Standard Error at national level for each Simulation Scenario

Scenario	Without Calibration	With Calibration
s01	0.001748757	0.00240419
s02	0.001718127	0.002759993
s03	0.001686163	0.001902013

For Scenario 1, the total relative bias is smaller (in absolute terms) when duplication calibration is included in the estimator (22) than when it is not (11) which mirrors the performance of this scenario across age-sex by local authority. However, for scenario 2 and 3, this does not hold and the total relative bias for these scenarios is larger (in absolute terms) when duplication calibration is included in the estimator (22). This is because for each domain (age-sex by local authority) the estimator with duplication calibration outperforms the estimator when there is no duplication calibration, however these estimates do not cancel each other out across the domains and cause the relative bias at national level to be negative. Moreover, for the relative root mean squared error and the relative standard error are larger across all the scenarios at national level for estimator (22) than (11). These results suggest that at national level the addition of duplication calibration to the estimator may not outperform the estimator where it is not included for given scenarios.

Summary

From the simulation study the inclusion of duplication calibration in the estimator (22) improves population estimates for each scenario across age-sex by local authority, where bias is smoothed out across these groups. Although at national level, for some scenarios the bias increases (in absolute terms) when estimator (22) is used, the largest relative bias value (scenario 2) is still very small. The estimator (22) does however increase the variance across these domains for each scenario, as expected from the inclusion of the wrong location model (15) and the duplication calibration ratio which is a ratio of the number of duplicates using both census-to-census and census-to-ccs linkage. It is expected the model selection methods that will be implemented will decrease this variance.

Questions for MARP:

1. Does MARP agree that the proposed duplication calibration approach should be used alongside the proposed estimation methods for the 2021 Census of England and Wales?
2. Any other feedback or comments.

6. Appendix

Scenario 1 where the undercoverage, overcoverage correct enumeration and wrong location model had the main effects of Region, HtC and Age-sex group. This shows how the estimators (11) and (22) perform when bias is not introduced into the simulation study through simple model selection for the correct enumeration model.

Figure 11: Relative Bias for Age-sex by Local Authority for Simulation Scenario 1 ordered by Age-sex by Local Authority

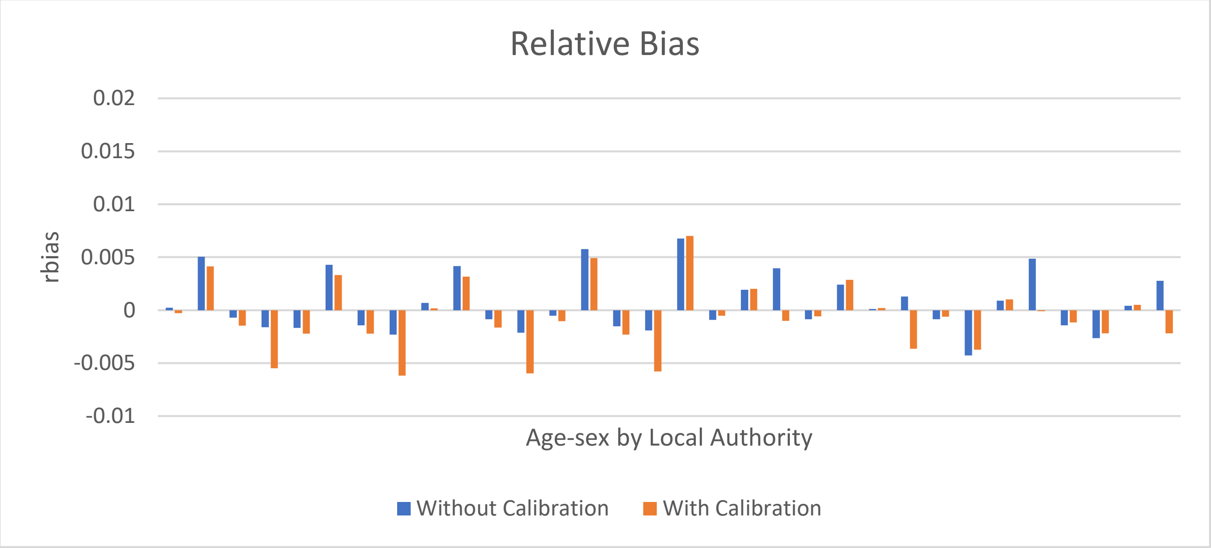


Figure 12: Relative Root Mean Squared Error for Age-sex by Local Authority for Simulation Scenario 1 ordered by Age-sex by Local Authority

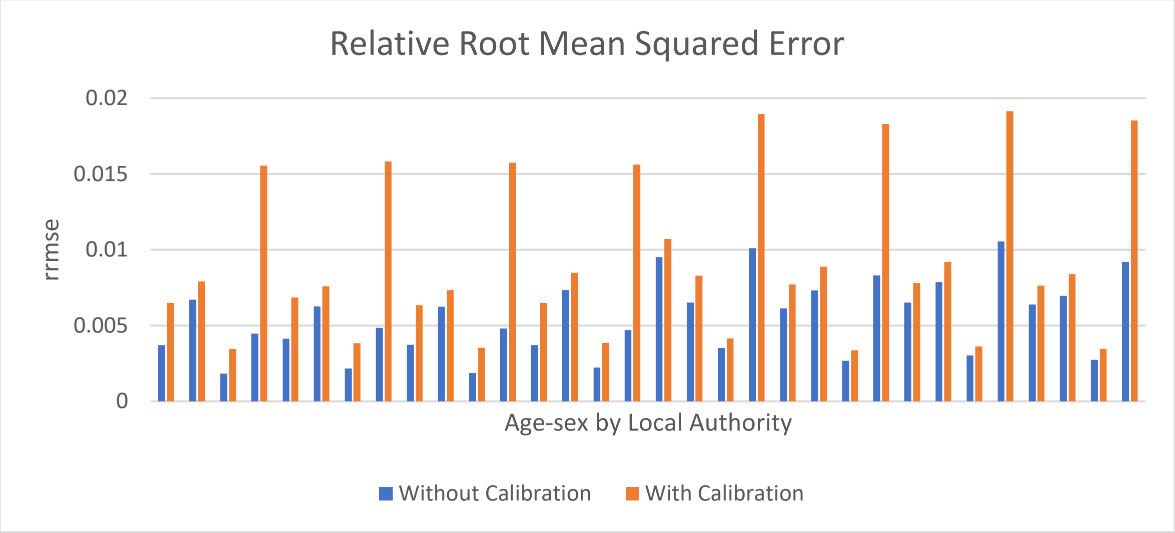
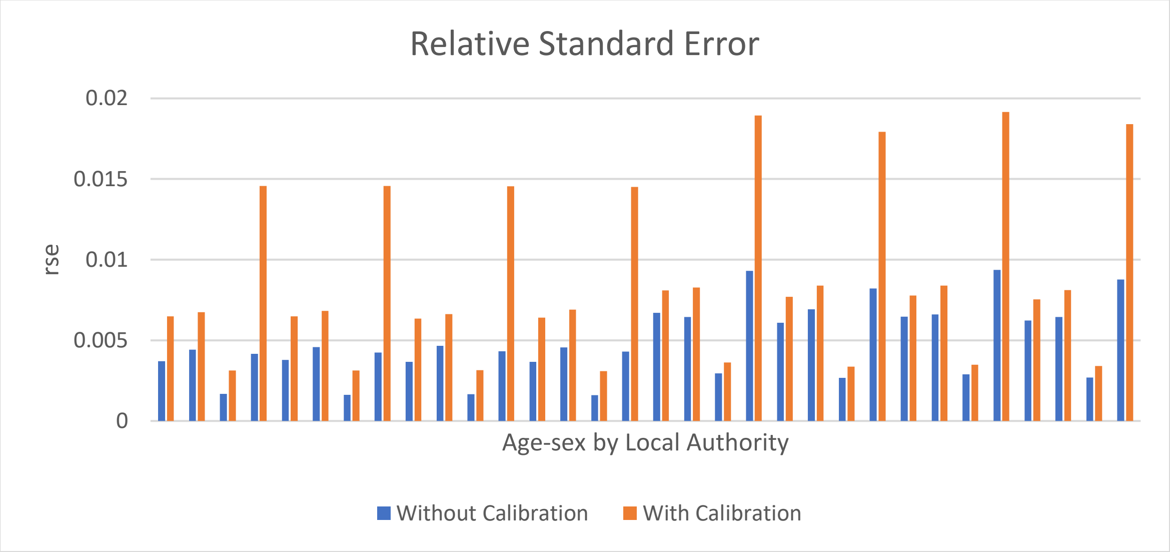


Figure 13: Relative Standard Error for Age-sex by Local Authority for Simulation Scenario 1 ordered by Age-sex by Local Authority



7. References

Abbott, O., and Brown, J., (2007) ‘Overcoverage in the 2011 UK Census’. Paper presented to 13th Meeting of the National Statistics Methodology Advisory Committee.

Abbott, O. and Large A. (2009) Measuring the level of duplicates in the 2011 Census. *Paper presented at 17th Meeting of the GSS Methodology Advisory Committee*.
Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/2008-2011/17th-meeting/gss-mac-seventeenth-meeting-booklet.pdf

Alho, J. (1990) Logistic Regression in Capture-Recapture Models. *Biometrics*, *46*, 623-635.

Alho, J., Mulry, M., Wurdeman, K. & Kim, J. (1993) Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. *Journal of the American Statistical Association, 88*, 1130-1136.

Brown, J. (2019) 'How can you use the duplicate search across census to combine with the over-coverage modelling?' *Method proposed after discussions on overcoverage estimation with Viktor Racinskij, Paul Smith, James Brown and Ceejay Hammond on 12th July 2019.*

Chambers, R. L. and Clark R. G. (2012) *An Introduction to Model-Based Survey Sampling with Applications*, Oxford University Press, New-York, USA.

Haldane, J.B.S (1945) On a method of estimating frequencies. Biometrika, Vol 33, No. 3, pp. 222-225

Large, A., Brown, J., Abbott, O. and Taylor, A. (2011) *Estimating and Correcting for Over-count in the 2011 Census.* ONS Survey Methodology Bulletin 69.
Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/smb-69/index.html

Office for National Statistics (2012), 2011 Census: Methods and Quality Report, *Overcount Estimation and Adjustment*. Available at:
https://webarchive.nationalarchives.gov.uk/20160108085304/http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf

Office for National Statistics (2014), 'Longitudinal Study 2011 Census Linkage Report', available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-userguide/quality-and-methods/quality/quality-assurance/index.html

Račinskij, V. (2018) Coverage Estimation Strategy for the 2021 Census of England and Wales. *Report presented at the Census External Assurance Panel on 16 October, 2018*. available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP105-Coverage-Estimation-Strategy-for-the-2021-Census-of-England-and-Wales.docx

Račinskij, V. & Hammond, C. (2019) Overcoverage estimation strategy for the 2021 Census of England & Wales, Office for National Statistics.

Shipsey, R. (2019) Methodology report on coverage matching for the 2021 Census, ONS internal report, available on request.

Shipsey, R. & White, Z. (2020) Census to census matching strategy 2021
https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP121-Census-to-census-matching-strategy-2021.docx

US Census Bureau (2008). 2010 Census Coverage Measurement Estimation Methodology. US Census Bureau, Washington, D.C. Available from
https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf

US Census Bureau (2012). 2010 Census Coverage Measurement Estimation Report: Aspects of Modeling. US Census Bureau, Washington, D.C. Available from
https://www.census.gov/coverage_measurement/pdfs/g10.pdf