Adjusting for the dependence bias in the coverage estimation of the 2021 Census of England & Wales

Viktor Račinskij

Office for National Statistics
Titchfield, UK

August 18, 2022
Draft version 0.4
Disclaimer: all results are preliminary and subject to revision

1 Introduction

Census coverage estimation produces the census coverage error adjusted population totals at the national and various small domain levels. These totals are often achieved by a combination of the survey sampling, capture-recapture, predictive modelling and small-area estimation methods, see for example Brown *et al.* (2019) and Baffour-Awuah *et al.* (2018). An overview of the coverage estimation approach proposed for the 2021 Census of England and Wales can be found in Račinskij (2018), Račinskij & Hammond (2019) and Račinskij (2020).

Inevitably, there is a number of conditions or statistical assumptions that must be met in order to guarantee approximately unbiased population size estimates. Particularly, the capture-recapture part of the coverage estimation is reliant on a large number of assumptions. To obtain the initial coverage error adjusted population totals for England & Wales, two data sources are used: the census data and the Census coverage survey. Hence, a special case of capture-recapture estimation, known as dual system estimation, is employed. In fact, it is a (mixed effects) logistic regression based version of the dual system estimator (Alho, 1990; Račinskij, 2018) that is proposed for the 2021 Census. The number of assumptions in capture-recapture methods for the census coverage models may vary depending on a model that is being considered (Wolter, 1986), but it may be argued that there are five main assumptions. Multiple captures occur in a closed population; different data sources can be perfectly linked; no spurious events such as counting in the wrong location or duplication are present; the capture probabilities are either constant (heterogeneous) at least in one of the sources or are uncorrelated between the sources; there is no casual dependence between the sources.

For every assumption listed above there are operational and statistical solutions that either guarantee that an assumption is satisfied or mitigate the effect of a departure

from the required condition. Even though the coverage survey takes place six weeks after the census day, the survey collects information on the usual residence at the census day; census to the coverage survey linkage is a combination of the automated and clerical review based decision process that has one of the highest practically attainable quality requirements; there is a process of estimating the overcoverage error (Račinskij & Hammond, 2019); the regression based approach ensures that the heterogeneity error is as small as possible (Alho *et al.*, 1993).

Dependence assumption, ways to minimize it and methods to adjust for dependence is a topic of this report.

2 Dependence

2.1 Overview

By independence in this report we mean the standard statistical definition of the term, that is the equality of the joint probability of a several events and the product of the corresponding marginal probabilities of each of the events. Thus, by dependence we mean that the joint probability cannot be factorized in such a way.

More specifically, in the context of two sources, we say that the census and coverage survey counts are independent if the joint probability, π_{11} that an element (individual or household) is counted in both census and the coverage survey equals to $\pi_{1+}\pi_{+1}$, where π_{1+} is the census inclusion probability, and π_{+1} is the coverage survey inclusion probability. Equivalently, independence in the dual system estimation holds whenever the cross product (odds ratio) of the cell probabilities (or cell counts $x_{ij}, i, j = \{0, 1\}$) equals to 1:

$$\theta = \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = \frac{x_{11}x_{00}}{x_{10}x_{01}} = 1,$$

where π_{10}, π_{01}, π_{00} are probabilities of being counted in census only, in the survey only and missed from both sources, respectively. Cell counts use the subscripts that have the same meaning as in the case of cell probabilities.

Whenever $\theta > 1$, there is a positive association between the census and survey counts. That is, an element that appeared (did not appear) in one of the sources is more likely to appear (to be missing) on the other source. Whenever, $\theta < 1$, there is a negative association between two counts, meaning that an element that appeared (did not appear) in one of the two sources is more likely to be missed (to appear) on the other.

Dependence between two sources incurs bias in the population size estimates. A positive association leads to a negative bias, while a negative association leads to a positive bias. Evidence from England & Wales as well as other countries suggest that it is positive association that occurs in the case of census and the coverage survey. Therefore, in practice any residual dependency is expected to manifest itself in underestimation of the population totals.

Since only two data sources are involved and the data are incomplete (missing in both sources in unobserved), there is no way of estimating the joint probability in the dual system framework other than as a product of the marginal probabilities. Thus,

there is no way of mitigating against dependence if dependence is present without some additional information.

2.2 Solutions: operational

We now will look at the ways of reducing the dependence between the sources and various ways of reflecting or adjusting for dependence.

Obviously, if it was possible to prevent inclusion in one of the sources to affect the inclusion in the other, this would result in meeting the condition of no dependence between the sources. Indeed, there is a number of procedures in place during the data collection that aim to minimise the dependence. For instance, the sampling frame for the coverage survey is as independent of the census address frame as possible. This results in the ultimate sampling units in the coverage survey to be postcodes and an independent address listing is carried out by the interviewers in the sampled postcodes prior to interviews. There are also restrictions on the extent of the census field interviewers' involvement in the coverage survey interviews.

While being indispensable, these procedures only protect from the operationally induced dependence and cannot deal with cases where positive association happens because of behavioral reasons.

High response rate provides protection against the dependence bias assuming that dependence does not get extremely high with the growing response rate. Response rate maximization in both census and coverage survey data is one of the key goals of the 2021 Census of England & Wales. In principle, putting more resources to obtain a very high response rate (say, 95%) across all the domains in the coverage survey might result in ignorable dependence effect. However, it is unlikely that uniformly high response rates across all the domains of interest can be achieved.

2.3 Solutions: theoretical

Capture-recapture methods are not limited to two data sources. An advantage of the multiple system estimation with k data sources is that it allows accounting for dependence between up to $k-1$ sources (Fienberg, 1972). Say, the triple system estimator with census, the coverage survey and some administrative list might be a natural candidate either to replace two source based estimators or to produce a set of alternative estimates that can be used to calibrate the main two source based estimates. For instance, the most complex model with three data sources could take into account (or test for) the case of a pairwise dependence between all sources, so that the cell probability $\pi_{ijk} = \alpha_{ij}\beta_{jk}\gamma_{ik}$, $i, j, k = \{0, 1\}$, for some probabilities $\alpha_{ij}, \beta_{jk}, \gamma_{ik}$ (not necessarily expressible by the marginal probabilities, see Agresti (2002)).

Solutions like triple system estimator are well justified from the basic theoretical standpoint. However, there is a number of practical issues that prevented the triple system estimator to be used in any of the censuses so far. The first difficulty is related to data linkage. Even linking two data sources to meet the quality standards for the coverage estimation is a huge operational undertaking. It takes many weeks and up to a

couple dozen of clerical matchers to complete the task. Linking three data sources with the required quality is a very difficult and time consuming process that might not be completed within the census time-frame and to the quality needed. The second difficulty is related to the properties of alternative data sources, such as administrative data. In fact, these data quite often do not satisfy various assumptions of the capture-recapture estimation (mentioned above). Notably, the assumptions of closed population and no spurious events such as overcoverage or erroneous inclusion (inclusion of elements that are not part of a target population, say, non-usual residents, individuals who left the country before the census day, etc.) do not always hold in administrative lists. As a consequence, while allowing to control for the dependence, the triple system estimation is likely to introduce additional bias. Some research shows that in certain situations for the average response rate achievable in census and the coverage survey, the simple dual system estimator has smaller mean square error than the triple system estimator even when dependence is present between the census and coverage survey data because of additional errors and the larger variance of the triple system estimator (Baffour-Awuah, 2009).

It is possible to adjust for the additional biases incurred by the third list. However, it will require additional stages of estimation that will bring additional variability into the final estimates. It may also require additional data collection since there is no easy way of detecting the erroneous enumerations. Some of the data collection needed to detect the erroneous enumerations in administrative data is not currently legally allowed. There exist attempts to use the latent class models to deal with erroneous enumerations within the triple system estimation. However, these methods may not be reliable enough to use them in the census coverage estimation.

2.4 Solutions: practical

Another family of methods that deals with the residual dependence bias in the coverage error corrected population size totals is focusing on adjusting the initial estimates using some external data (Bell , 1993). These data could, for instance, be demographic analysis based totals, sex ratios, etc. or some alternative data based estimates of some units of interest. In this case no additional linkage is required and some reasonably reliable alternative sources may be available. Variants of this approach were used in 2001 and 2011 Censuses of England & Wales to adjust for the dependence bias.

In both 2001 and 2011 Censuses the alternative estimate of the occupied households, known as an alternative household estimate, was used as external data set for bias adjustment. The alternative household estimate is an estimate based on the combination of the valid census returns for the usual residents and completed census dummy forms for non-responding households, see ONS (2012) for more details. The alternative household estimates are produced for the coverage survey sampled areas at some level of aggregation like an estimation area by hard-to-count, for instance.

Among advantages of the alternative household estimate is that it is based on the data collected by the Office for National Statistics and so there is a full control over the data collection and delivery time. The data collection is also a part of the general

collection operations (say, dummy forms are used to produce statistics on the number of second homes). Furthermore, this approach has already been used twice.

Among limitations and challenges of this approach one can mention some non-trivial work of obtaining these estimates due to very limited information available when completing a dummy form. So a careful, conservative and robust approach is required when deriving the alternative estimates since, as it will be shown later, households estimates are going to be calibrated to the alternative household estimates at the levels of alternative estimate post-strata. The alternative estimates are available for the household population only, so the individual population adjustment is always based on some indirect synthetic method. Also, the alternative estimate is available only for post-strata formed crossing a geography (like local authority) by hard-to-count by accommodation type (terraced, detached, etc.). Therefore, there is a high level of syntheticity involved here (it is possible to obtain estimates at, say, local authority by hard-to-count by output area level and then fit an area level model, but some work done showed that there was no much gain over the approaches presented later in this paper).

3 Bias adjustment in 2001 and 2011 Censuses of England & Wales

Methodology behind the dependence bias adjustment in the 2001 and 2011 Censuses of England & Wales is summarized in Brown *et al.* (2006). In both cases the alternative household estimate was used to estimate the household level odds ratios first and then (assuming that several assumptions hold) derive an individual level synthetic estimate.

The dependence bias adjustment approximately added 230,000 (slightly less then 0.5% of the overall population) and 584,000 (around 1% of the overall population) individuals to the overall population total in 2001 and 2011 Censuses, respectively. Note, that in two previous censuses, the dependence bias adjustment also corrected for the residual heterogeneity bias. The rough estimate of the heterogeneity bias for the 2011 Census relative to the population total is 0.22%. It is expected that the heterogeneity will be dealt better in 2021 Census because of the regression based estimation.

4 Bias adjustment in 2021 Census

4.1 An overview

It is proposed to obtain the alternative household estimate in a similar way to how it was done in two previous censuses. However, the method described in Brown *et al.* (2006) is not applicable for the regression based estimation. Therefore, a number of methods were developed to allow the dependence bias adjustment of the logistic regression based household and individual population totals.

4.2 Recap on the coverage estimation in 2021 Census

The general framework for the census coverage error corrected population size estimation is based on the mixed effects logistic models. Here for simplicity we ignore the overcov-

erage estimation and provide a quick summary of obtaining the undercoverage adjusted estimates. We will refer to the estimates that are not bias adjusted for dependence as initial estimates.

We are interested in the population total for the domain vL, which can be estimated using the following mixed effects logistic regression based estimator:

$$\hat{T}_{vL} = \sum_{r \in vL} \hat{\pi}_r^{-1} = \sum_{r \in vL} \left[\frac{1}{1 + \exp\left(- \left[\mathbf{x}_r^T \hat{\boldsymbol{\beta}} + \mathbf{z}_r^T \hat{\boldsymbol{\kappa}} + \hat{u}_L \right] \right)} \right]^{-1} \tag{1}$$

Where: the probability π_r that an element (individual or household) r is captured in census; v – a covariate or combination of covariates (say, age-sex, tenure, age-sex by tenure, etc.) of interest; L – local authority of interest; $\mathbf{x_r}^T$ – vector of main effects and interactions based on census / coverage survey covariates (such as age-sex, tenure, ethnicity, household size, etc.); $\mathbf{z_r}^T$ – vector of main effects and interactions based on design variables and field management information (hard-to-count index, observed census return rate at the local super output area; $\hat{u}_L$ – random local authority effect.

The household population estimator is conceptually very similar to the individual population estimator. The main difference is in the set of variables used as individual model allows using directly both individual and household variables, while household model can only directly use household variables and indirectly individual variables (via the derived variables like household structure that combines the age-sex, marital status and relationship variables).

4.3 Adjusting household population estimates

The method for adjusting the household estimates is first discussed (this is going to be our default household population adjustment method).

We carry on the estimation process as outlined in the previous section by fitting the mixed effects logistic regression to estimate the initial census household response probability of a record r having a covariate pattern $\mathbf{y}$:

$$\hat{\tau}_r = \frac{1}{1 + \exp\left(- \left[\mathbf{y}_r^T \hat{\boldsymbol{\beta}} + \mathbf{z}_r^T \hat{\boldsymbol{\kappa}} + \hat{u}_L \right] \right)} \tag{2}$$

These probabilities can be used to produce the initial (undercoverage) error corrected population totals for a domain of interest. If the dependence is present, however, those estimates will be negatively biased.

To deal with the dependence bias we first obtain the alternative household estimates $\hat{T}_{Lht}^{(alt,hh)}$ at the alternative household post-stratum level formed by local authority (L), hard-to-count index (h) and accommodation type (t) (we also allow for collapsing of t within a hard-to-count to prevent any unstable estimates)

Since the estimator 1 allows producing estimates virtually for any domains, we produce the dependence bias unadjusted household estimates for the alternative household

post-strata:

$$\hat{T}_{Lht}^{(hh*)} = \sum_{r \in Lht} \hat{\tau}_r^{-1}.$$

We can now obtain the dependence bias adjustment weights for the alternative household post-strata:

$$\hat{w}_{Lht}^{(hh)} = \frac{\hat{T}_{Lht}^{(alt,hh)}}{\hat{T}_{Lht}^{(hh*)}} \tag{3}$$

Finally, we use the dependence adjustment weight to obtain a dependence bias adjusted estimate for a domain of interest

$$\hat{T}_{bL}^{(hh)} = \sum_{r \in bL} \sum_{r \in ht} \hat{w}_{Lht}^{(hh)} \hat{\tau}_r^{-1}.$$

This notation reflects the fact that often a domain bL is split across multiple alternative household post-strata (say, tenure 'owns with mortgage or loan' in local authority L is split across all hard-to-count and accommodation type levels) and that local authority, hard-to-count index and accommodation type are matched between the initial non-response and adjustment weights. This notation is cumbersome, and instead of it in what follows we simply write

$$\hat{T}_{bL}^{(hh)} = \sum_{r \in bL} \hat{w}_{Lht}^{(hh)} \hat{\tau}_r^{-1}.$$

There is a number of important assumptions behind such estimation. First, we assume that the alternative household estimates are of very high quality, any errors are ignorable. It is important since $\hat{T}_{bL=Lht}^{(hh)} = \hat{T}_{Lht}^{(alt,hh)}$. In other words, at the alternative household post-strata level the bias adjusted household estimates are equal to the alternative household estimates. Another important assumption is the extent of dependence is uniform within each alternative household post-stratum, so that $\hat{w}_{Lht}^{(hh)}$ is applicable for all $r \in Lht$.

4.4 Adjusting individual population estimates: a direct method

Two methods for adjusting individual population estimates were developed and tested so far. The first one is called the direct method and combines the bias adjustment weights 3 from the household adjustment stage with the initial non-response weights for individuals:

$$\hat{T}_{vL}^{(hh)} = \sum_{r \in vL} \hat{w}_{Lht}^{(hh)} \hat{\pi}_r^{-1}. \tag{4}$$

In addition to all mentioned assumptions, this estimator assumes absence of within household dependence and that an effect that the dependence has on the household units is the same as for the individual units.

4.5 Adjusting individual population estimates: within household assisted method

The second method is slightly more involved and requires additional modelling effort to estimate the probability of individual response within responding households. This method may have a better justification from the theoretical point of view and, if the alternative household data were at the individual level and as rich in variables as the original census data, would arguably give the 'best' estimates. But due to the fact that the alternative household data are very chunky, it may not perform as good as expected.

We start working out the alternative household response probabilities using the observed census data and the alternative household estimates at the Lht level, simplistically:

$$\hat{\tau}_{Lht}^{(alt)} = \frac{x_{Lht}}{\hat{T}_{Lht}}, \tag{5}$$

where x_{Lht} is the observed census count.

Using the chain rule we have

$$P(\text{individual responds}, \text{household responds}) =$$
$$P(\text{individual responds} \mid \text{household responds}) P(\text{household responds}).$$

The alternative (dependence unaffected probability) of household response at the alternative household post-stratum level is 5. It is possible to model the response probability for individuals in the responding households. Use the subpopulation of all households of the size >= 2 that are present in both the census and the Coverage survey, and estimate the within household response probabilities $\hat{\pi}_r^{(whh)}$ using the logistic regression.

Once within household response probabilities are estimated, we can estimate the joint response probabilities

$$\hat{\pi}_r^{(adj)} = \begin{cases} \hat{\pi}_r^{(whh)} \hat{\tau}_{Lht}^{(alt)}, \text{if household size >= 2} \\ \hat{\tau}_{Lht}^{(alt)}, \text{otherwise} \end{cases} \tag{6}$$

Note that that local authority, hard-to-count index and accommodation type are matched between the initial non-response and the alternative household probabilities (just as weights in case of the direct method).

Now we can work out the adjustment weight at some level (usually just Lht)

$$\hat{w}_{Lht}^{(p)} = \frac{\sum_{r \in Lht} (\hat{\pi}_r^{(adj)})^{-1}}{\sum_{r \in Lht} \hat{\pi}_r^{-1}} \tag{7}$$

Finally, we estimate for any domain of interest as

$$\hat{T}_{vL}^{(adj)} = \sum_{r \in vL} \hat{w}_{Lht}^{(p)} \hat{\pi}_r^{-1} \tag{8}$$

5 Simulation study

We conducted a simulation study to explore the performance of the above estimators. These simulations built on the Brown and Sexton (2009) and specifically on Račinskij (2018) and Račinskij (2019), but allow dependence between census and the coverage survey. In these simulations 128 instances (iterations) of census and the coverage survey are generated for entire England &Wales for each scenario (described below) from the mixed effects logistic models fitted to the 2011 census coverage survey cluster data (linked census to the survey data). The vector of covariates $\boldsymbol{x}_i$ in the census model includes continuous age (modelled using the natural cubic splines), activity last week, accommodation type, address one year ago, born in the UK indicator, hard-to-count, household relation, household size, marital status, ethnicity, region, self-contained accommodation indicator, sex, short-term migrant indicator, tenure and various interactions of the above variables, see Račinskij (2019).

Parameters for the dependence are based on the estimated odds from the 2011 Census. These are synthetic estimates at the estimation area by hard-to-count by age-sex group level. These odds can be adjusted to simulate different levels of dependence.

There are three simulation scenarios, with the census and coverage survey non-response patterns being similar to those observed in 2011 (with 94% overall census response, etc.), the patterns are the same across all scenarios, but the level of dependence between the two sources varies. The first scenario is a benchmark and has no dependence between the data sources. The second scenario has dependence that results in the underestimation of the population total at the national level by 0.5%, this scenario is (arbitrary) referred to as the low dependence scenario. In the third scenario the level of dependence results in the underestimation of 1.05%, this scenario is referred to as the medium dependence scenario. Thus, second and third scenarios roughly correspond to the estimated combined effects of the dependence and heterogeneity in 2001 and 2011 Censuses, respectively. The level of dependence in 2011 Census without the heterogeneity is thus somewhere between the levels of two dependence scenarios considered in the study.

In terms of the models used in the estimation, the household model is the mixed effects logistic model with random local authority effect; main effects include household structure, household size, accommodation type, tenure, hard-to-count index, household ethnicity, hard-to-count score (continuous), region; region and household ethnicity; interactions include region with hard-to-count score and region with tenure. The individual model is the mixed effects logistic model with random local authority effect; main effects include age-sex, tenure, ethnicity, accommodation type, household size, marital status, relationship, address one year ago, activity last week, hard-to-count index, hard-to-count score (continuous); interactions include region with ethnicity, region with activity last week. These models are not 'optimal' ones, but give a reasonable balance between the quality of estimates and running time.

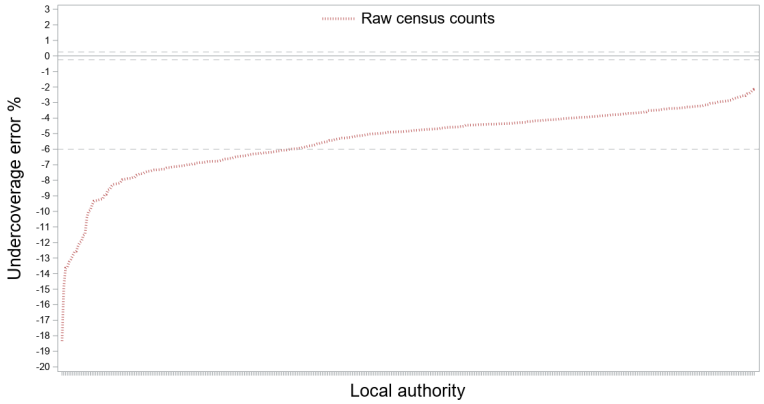

Figure 1: Raw census count by local authority

6 Results

As a recap, we present the chart showing what the raw census counts look like at the local authority level (Figure 1). We start with the results for the low dependence scenario with no dependence case for the benchmark, no bias adjustment estimation and the direct method 4. We focus on the person population. In terms of relative bias for the age-sex totals (Figure 2), the benchmark scenario produces good quality estimates with some peaks / dips for some young adults groups mainly due to some model misspecification and residual heterogeneity (not presented here, but which is notably smaller than in the case of the dual system / ration / synthetic approach). When the dependence is present and no adjustment for it is done, the relative bias is at least -0.25% for all age-sex groups and can be as large as -1.0%. The direct adjustment method 4 results in less biased estimates compared to the no adjustment method (no more than +/- 0.25% for most of the groups), but, is not unbiased. We will discuss the reasons for that in the next section. The adjustment tends to over-adjust those 50+ age-sex group and slightly under-adjust young adults.

As for the relative root mean square error (Figure 3), it is clear that the direct adjustment results in lower error than the unadjusted estimates (the lower the better), except for some 60 to 75 year old groups, where adjustment may result in a higher error than in unadjusted estimates. Note, that in some cases the relative root mean square error of the adjusted estimates may be even lower than that of the benchmark scenario (young males). This is because the alternative household estimates are perfect in our simulation study for each coverage sample drawn and thus provide variance reduction in the final estimates.

Figure 4 shows the relative bias at the local authority level (results are sorted by the relative bias for the direct adjustment method, from the smallest to largest). Results for some of the ‘rightmost‘ local authorities can be ignored as those are very small local authorities that have an unstable adjustment. In live census run, they should be collapsed with the neighboring areas at the adjustment stage. It can be seen, that for the

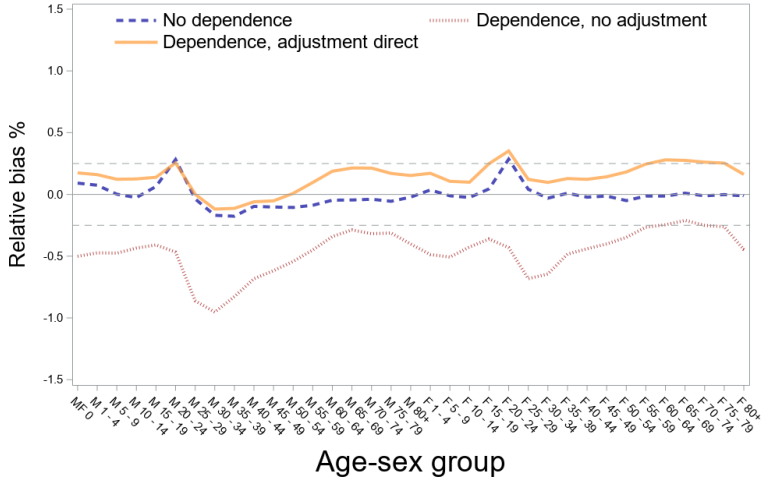

Figure 2: Relative bias, age-sex totals

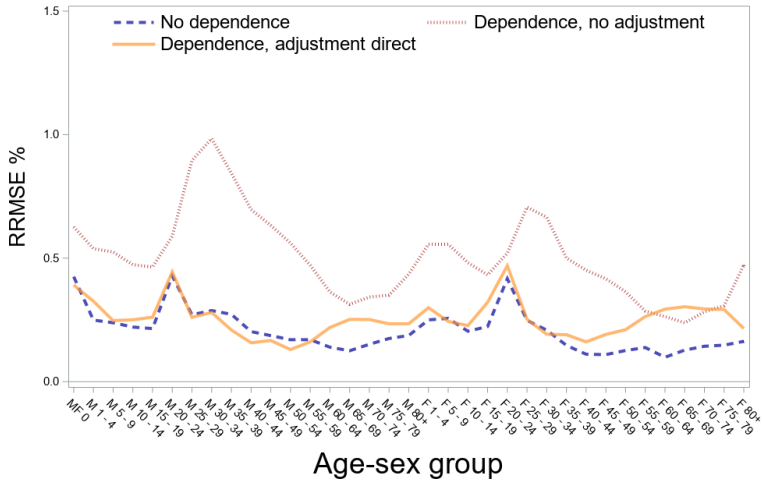

Figure 3: RRMSE, age-sex totals

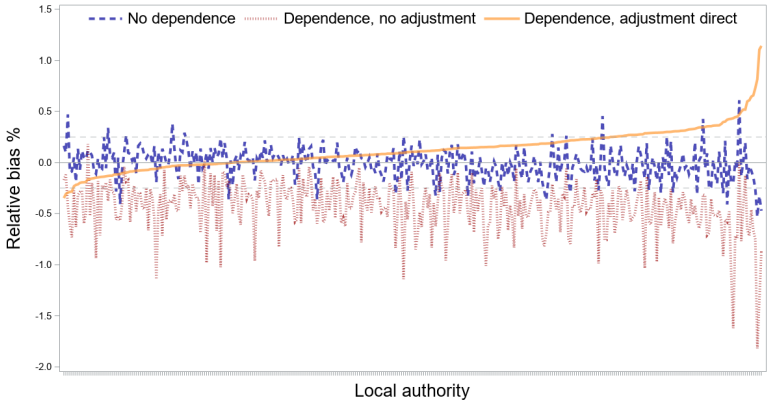

Figure 4: Relative bias, local authority totals

Figure 5: RRMSE, local authority totals

majority of local authorities the direct adjustment reduces the relative bias. However, when looking at the relative root means square error (Figure 5) for the local authority estimates, it is obvious there is a substantial number of areas for which overall relative error for the adjusted estimates is higher than for the unadjusted ones. In other words, while the adjustment makes age-sex totals to be closer to the unknown population totals, it may contribute to an additional error in the local authority estimates.

Frequently, there are some reasonably reliable benchmarks for age-sex groups at the national level (like demographic sex ratios) while it is unlikely to have anything better than the census figures for local authorities. Hence, one can expect that whatever level of the dependence is, the decision to make an adjustment will be based on the age-sex national totals.

We now move to the medium dependence case. The estimation scenarios are as before, but we are also testing the performance of within household assisted method 8. Results for the relative bias are presented in (Figure 6). The pattern is similar to the

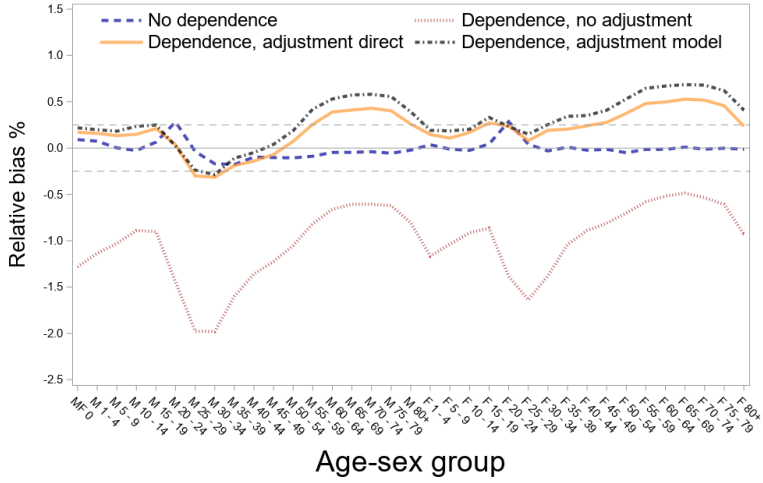

Figure 6: Relative bias, ags totals

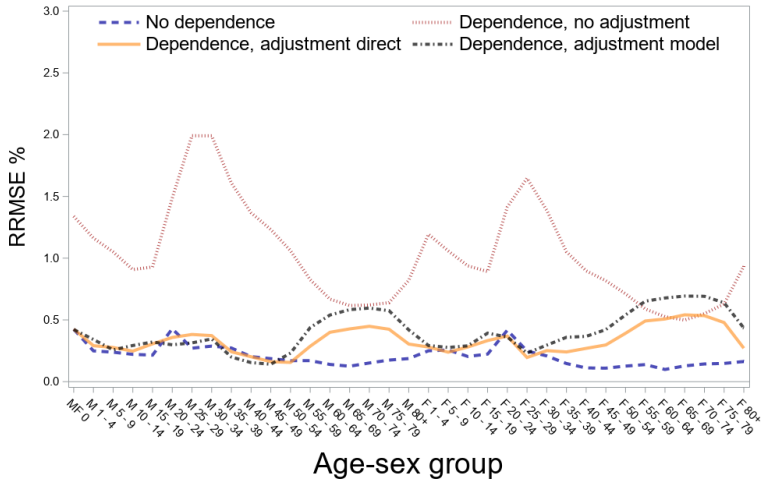

Figure 7: RRMSE, ags totals

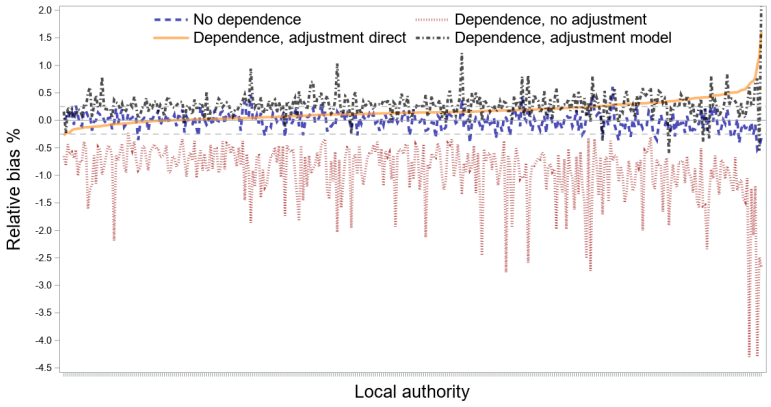

Figure 8: Relative bias, local authority totals

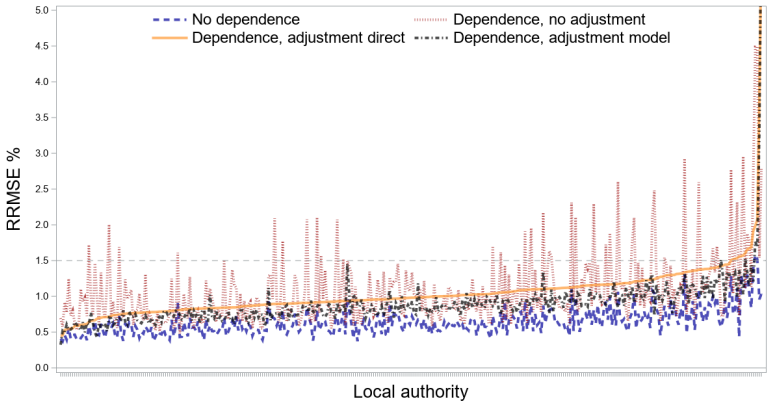

Figure 9: RRMSE, local authority totals

one already seen for the low level of dependence. The relative bias is quite large when adjustment is absent. Both adjustment methods result in bias reduction, with underestimation and overestimation of younger and older age-sex groups, respectively. The within household assisted method tends to overestimate more than the direct method.

The relative root mean square error (Figure 7) demonstrates how large is error reduction for those 0 to 60 years old, but some of 60 to 75 are sometimes doing worse with the adjustment than without it. This issue is discussed and a number of solutions is proposed in the next section.

As for the local authority estimates (Figures 8 & 9), it is clear that in the medium dependence case, bias adjustment pays off for the majority of local authorities.

All the results for the age-sex by local authority are presented in the Appendix.

7 Improving the methods

7.1 Apportionment based on the original census non-response weights

In all the approaches outlined above, there is a single adjustment weight for all census records within an alternative household stratum. Since the alternative household stratum is formed by a local authority, hard-to-count and accommodation type, it is expected that elements with varying non-response and dependence propensities are going to be pooled together. This in turn may result in underadjustment for one part of the the stratum population and overadjustment for another. In fact, this behaviour is present in the estimation work presented, with the estimates for several 50+ age groups being overadjusted. This can be explained by the fact that individuals in these groups tend to have higher response rate and lower dependence propencities, but reside in areas with individuals who have higher dependence propensities and who have disproportional contribution to the resulting dependence error. To mitigate for this effect, an apportionment (re-weighting) approach is proposed.

This approach is based on the estimated census non-response weights. There is a number of considerations related to such apportionment. First, any apportionment based on the census non-response weights is biased since the original non-response weights are subject to the dependence bias. Nevertheless, if a contribution of a certain group within an alternative household stratum to the adjustment weight is correlated to the census non-response, we also expect correlation to hold with the estimated non-response weight, unless the dependence is extreme.

The second consideration is related to the fact that the error incurred by dependence between the two sources is a function of three parameters: the survey response probability, the census response probability and the odds ratio between the cell probabilities. It implies that an accurate apportionment may require the estimated coverage survey response probabilities to be taken into account. It is possible to argue, however, that the estimated census coverage probabilities will be correlated with the survey coverage probabilities (this correlation should not be confused with the correlation that results in the heterogeneity bias) and therefore using only the estimated census coverage weights is

sufficient for the apportionment. Indeed, the Pearson correlation coefficient between the (true) coverage probabilities of the two sources ranges between 0.48 and 0.75 for 80% of the alternative household strata. If it is not sufficient, a relatively easy way of taking the survey response into account is to fit the census coverage model and estimate the survey coverage probabilities. These probabilities then can be used alongside the estimated census coverage probabilities to post-stratify the alternative household estimates in a way that will reflect the joint coverage probability. The apportionment method based on the census non-response weights itself would stay unchanged.

We can now discuss the general approach to apportionment method for the household and person populations. For the household population, we start with 3 as before. The apportionment is done using an apportionment variable. Throughout this work, the household structure is used as apportionment variable. This variable reflects the broad age-sex grouping, broad household size grouping (household of size 1, and the rest) and relationship. The advantage of this variable is that it fuses individual and household level information, which may prove useful given the fact that the alternative household estimates are available only for the household population. The disadvantage of this variable is that it is subject to data collection errors. Say, if within household non-response occurs in a responding household of size 2, the household will be incorrectly assigned structure of single person household.

We use the estimated census probabilities to work out the weighted mean of the household non-response weights for the alternative household stratum plus the apportionment variable a:

$$\hat{\tilde{w}}^{(hh*)}_{Lhta} = \frac{\sum_{r \in Lhta} \hat{\tau}_r^{-1}}{n_{Lhta}}, \tag{9}$$

where n_{Lhta} is the number of census cases that belong to $Lhta$. Note that there is an implicit weighting within $Lhta$ by the characteristics not reflected by the stratum itself, but pooled within. Say, if the household structure is 'related male and female aged 20 - 34 with children', the above weight will reflect the observed census frequencies of all age-sex groups that happened to belong to this household structure.

We also obtain the weighted mean of the household non-response weights for the alternative household stratum:

$$\hat{\tilde{w}}^{(hh*)}_{Lht} = \frac{\sum_{r \in Lht} \hat{\tau}_r^{-1}}{n_{Lht}}. \tag{10}$$

We can now estimate the dependence bias weight for the household population not only at the Lht, but for also at $Lhta$ stratum:

$$\hat{w}^{hh}_{Lhta} = 1 + (1 - \hat{w}^{hh}_{Lht}) \frac{\hat{\tilde{w}}^{(hh*)}_{Lhta} - 1}{\hat{\tilde{w}}^{(hh*)}_{Lht} - 1}, \tag{11}$$

provided $\hat{w}^{hh}_{Lht} > 1$ (in real implementation it's a bit more involved and depends on how we want to treat the cases where the alternative household estimate is smaller than the original one: keep calibrating, or use the original weight).

The rest is similar to the estimation approaches as already presented. For the household population, we estimate a domain of interest using the following estimator:

$$\hat{T}_{bL}^{(hha)} = \sum_{r \in bL} \hat{w}_{Lhta}^{(hh)} \hat{\tau}_r^{-1}.$$

The apportioned version of the direct estimator for individuals is then

$$\hat{T}_{vL}^{(hha)} = \sum_{r \in vL} \hat{w}_{Lhta}^{(hh)} \hat{\pi}_r^{-1}. \tag{12}$$

For within household assisted method we obtain the alternative household response probability at $Lhta$ level:

$$\hat{\tau}_{Lht}^{(alta)} = \frac{x_{Lht}}{\hat{T}_{Lhta}}, \tag{13}$$

where $\hat{T}_{Lhta}$ is dependence bias corrected estimate for the number of households belonging to $Lhta$ (this is different from how 5 is computed). From here, we proceed in a familiar way with appropriate weights.

Estimate the joint response probabilities

$$\hat{\pi}_r^{(adja)} = \begin{cases} \hat{\pi}_r^{(whh)} \hat{\tau}_{Lhta}^{(alt)}, \text{if household size} >= 2 \\ \hat{\tau}_{Lhta}^{(alt)}, \text{otherwise} \end{cases} \tag{14}$$

Work out the adjustment weight at $Lhta$:

$$\hat{w}_{Lhta}^{(p)} = \frac{\sum_{r \in Lhta} (\hat{\pi}_r^{(adja)})^{-1}}{\sum_{r \in Lhta} \hat{\pi}_r^{-1}} \tag{15}$$

Estimate for any domain of interest

$$\hat{T}_{vL}^{(adja)} = \sum_{r \in vL} \hat{w}_{Lhta}^{(p)} \hat{\pi}_r^{-1}. \tag{16}$$

If we want to reflect the fact that dependence error also depends on the survey response probability, we can work as follow. In addition to all the modelling mentioned above, fit a model to estimate the Census coverage survey response probabilities. Within sampled areas use these and census coverage probabilities to compute the joint response probabilities. Compute means of the joint probabilities for each sampled output area. Work out the quartiles of the output area means for each hard-to-count by region. Post-stratify output areas within hard-to-count by region into four strata as follows: 1 – if the mean of the joint probabilities for an output are is below Q1, 2 – if it is between Q1 and Q2, 3 – if it is between Q2 and Q3, 4 – otherwise. We call this hard-to-count by region specific strata hard-to-count tiers. When computing the alternative household estimate, include this tiers in stratification, so that the alternative household estimates are obtained for a local authority by hard-to-count by hard-to-count tier by broad accommodation type (3 levels only in this work). We collapse accommodation

type within a tier further if needed. All estimator remain as described above, only the post-strata are different. It can be argued that the survey model needs not to be as good as the census model as it is only required for the post-stratification. In this work we used the true joint probabilities for the proof of concept and to save some time on additional estimation components (the survey model).

7.2 Results

We look at the estimation result with and without the apportionment. A few notes of caution need to be made. First, when dependence is present and the bias adjustment weights are estimated and incorporated into the estimation, it becomes extremely difficult to disentangle the residual dependence error from all the remaining sources of errors. Second, due to the fact that estimation for the person population is relatively time consuming, we run fewer estimation iteration (128) for each scenario than is desirable. Therefore, some of the observed results may be attributed for the insufficient number of runs (say, for the default method in the case of household population we would expect the relative bias to be closer to 0). Finally, as the number of scenarios to consider grows rapidly, we had to present just a selection of the scenarios which is not free of some subjective choice.

With all the above caveats, it can seen in Table 1 that doing the apportionment or apportionment with the hard-to-count tiers in general leads to improved results at the national level. In terms of the relative bias for the age-sex totals, both the apportionment and apportionment with the tiers help to resolve the unwanted overadjustment for 50+ age-sex groups. However, since the apportionment method relies on the non-response weights from the main estimation model, whenever some sort of model misspecification is present, it can translate to the bias adjustment. The case of males and females aged 20-24 is a good example. From the results with no dependence it can be seen that the census coverage model is not quite well specified for these age-sex groups. This misspecification translates to the overadjustment of these groups. Obviously, one of the most important goals of the coverage estimation is to get the census coverage model as good as possible and avoid substantial misspecification.

Regarding the relative root means square error for the age-sex totals, an improvement for the vast majority of domains can be seen, with noticeable exception for the two groups for which the model is misspecified.

Apportionment (with and without the hard-to-count tiers) also in general results in bias reduction for the local authority totals. However, using the apportionment methods as they are set up in this study may introduce additional variability sometimes resulting in larger variance than the approaches without apportionment. We can argue that the variance can be reduced by more accurate choice of number of tiers for each region (with possible split between rural and urban area within a region) and smarter collapsing of the accommodation type within the tiers.

Population	Scenario	Adjustment method	RB%	RSE%	RRMSE%
Household	No dependence	NA	-0.055	0.046	0.072
	Low	No adjustment	-0.616	0.039	0.618
	Low	Default	0.045	0.042	0.062
	Low	Apportionment	0.063	0.042	0.074
	Low	Apportionment with tiers	-0.004	0.046	0.046
	Medium	No adjustment	-1.294	0.038	1.295
	Medium	Default	0.033	0.044	0.055
	Medium	Apportionment with tiers	-0.084	0.047	0.097
Person	No dependence	NA	-0.009	0.050	0.051
	Low	No adjustment	-0.483	0.052	0.486
	Low	Direct	0.130	0.047	0.139
	Low	Direct apportionment	0.080	0.049	0.094
	Low	Within hh LR apportionment	-0.043	0.150	0.156
	Low	Direct apportionment with tiers	0.026	0.055	0.061
	Medium	No adjustment	-1.053	0.047	1.054
	Medium	Direct	0.179	0.056	0.188
	Medium	Within hh LR	0.276	0.048	0.280
	Medium	Direct apportionment with tiers	0.008	0.061	0.062

Table 1: Quality of the estimates at the national level

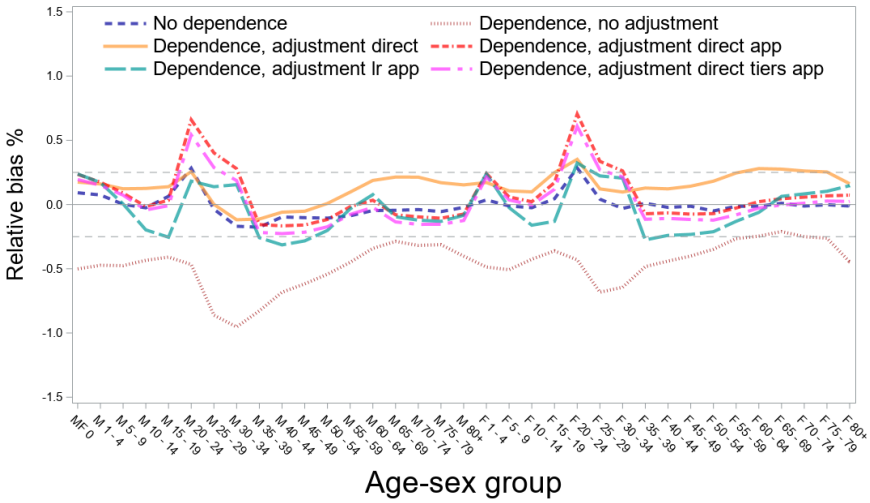

Figure 10: Relative bias, age-sex totals, low dependence ('lr' stands for the within household logistic regression model, 'app' stands for apportionment

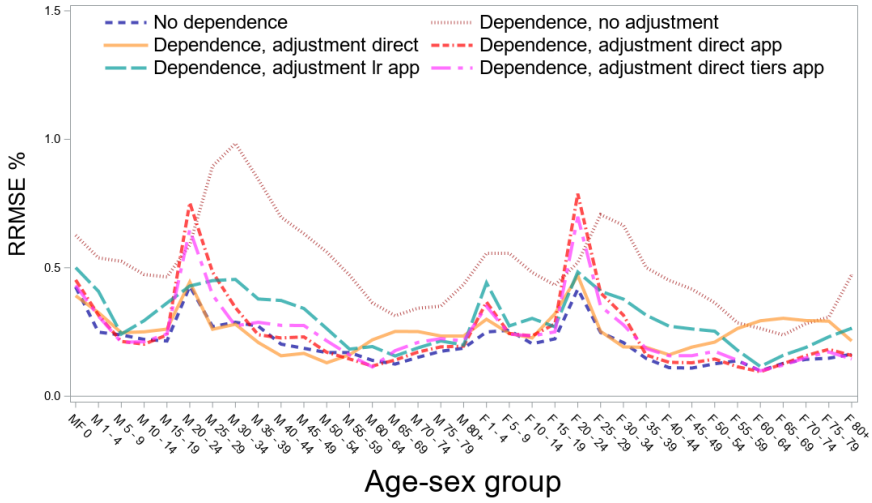

Figure 11: Relative root mean square error, age-sex totals, low dependence

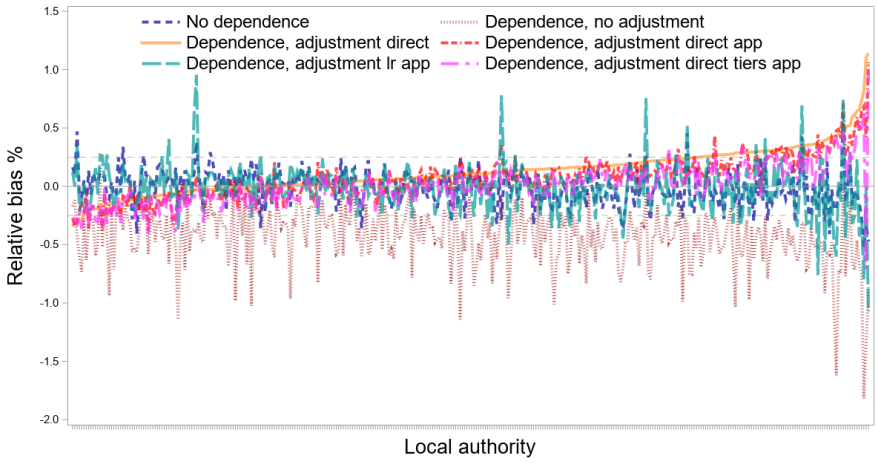

Figure 12: Relative bias, local authority totals, low dependence

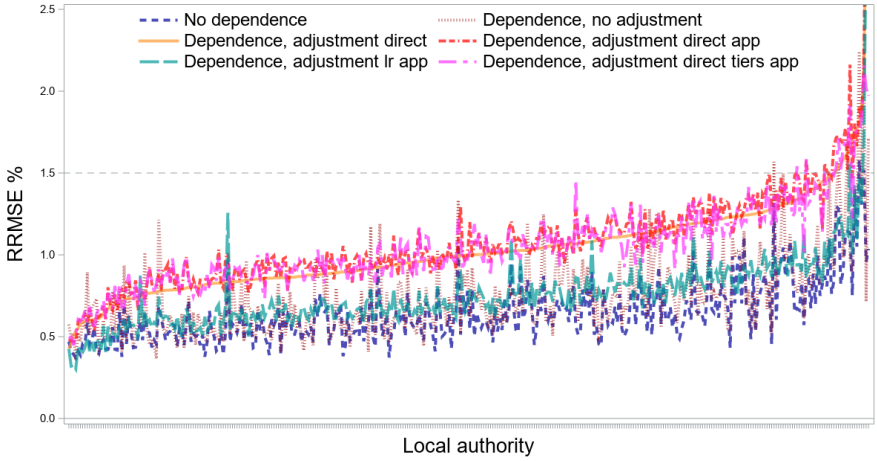

Figure 13: Relative root mean square error, local authority totals, low dependence

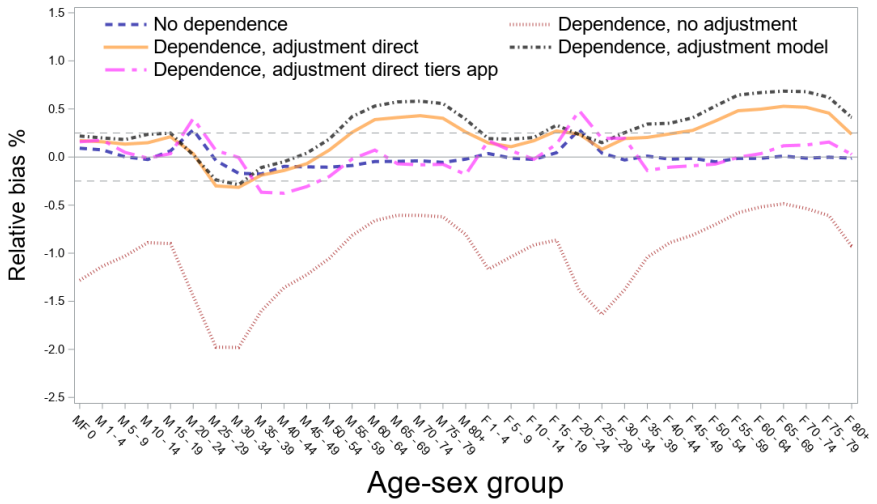

Figure 14: Relative bias, age-sex totals, medium dependence

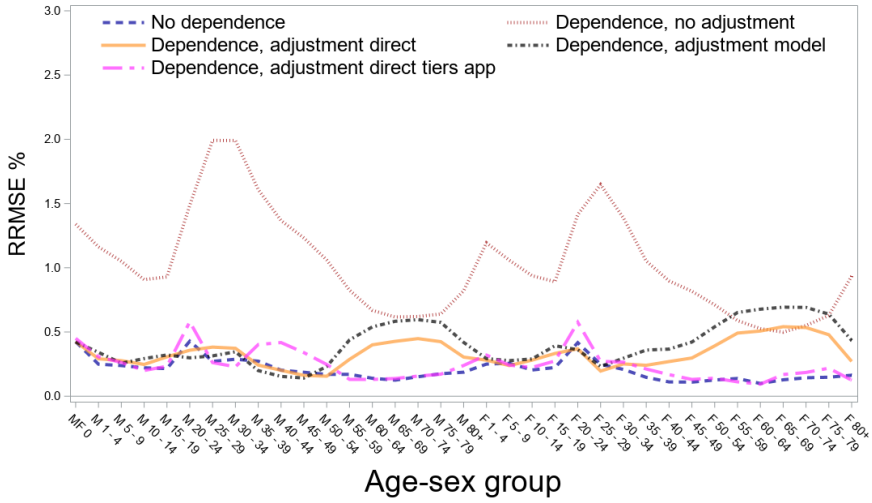

Figure 15: Relative root mean square error, age-sex totals, medium dependence

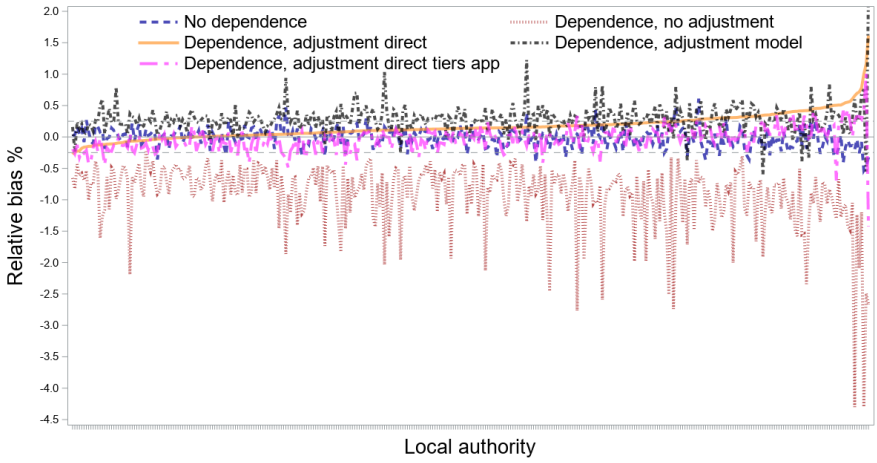

Figure 16: Relative bias, local authority totals, medium dependence

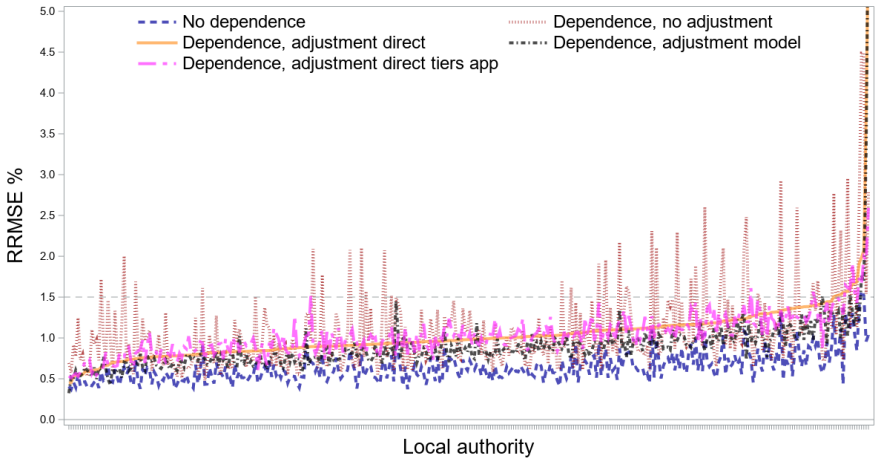

Figure 17: Relative root mean square error, local authority totals, medium dependence

8 Discussion, future work and recommendations

In this report the initial work on several dependence bias adjustment approaches for the logistic regression based coverage error corrected population totals were presented. It was demonstrated that these methods result in bias reduction both for the age-sex and local authority totals. Bias adjustment discussed here also reduces the relative root mean square error for the age-sex totals, but reduction for the local authority totals depends on the strengths of association between the coverage survey and census.

It is clear that none of the methods presented is capable of producing unbiased estimates. There are multiple reasons for that. All methods considered adjust person population indirectly using the information from the household population. This introduces several discrepancies. As an extreme example, imagine that only households of size 1 are subject to dependence. The alternative household estimate is capable in theory to get a correct number of *all* households within a post-stratum and so a correct adjustment weight for all household in the post-stratum is computed. Then the weight is applied for all observed persons within the post-stratum, which includes all individuals in households of size greater than one and results in the overestimation of the post-stratum total. Another reason is that the alternative household estimate can be derived only at a high level and thus pools together households and individuals with very different characteristics, non-response and association propensities. The apportionment method presented above partly mitigates the issue of pooling, but the price is larger variability at lower levels of geography.

As for the within household assisted approach, an additional reason to those mentioned is current model simplicity. There is some additional research work going to establish the upper bound of performance for this method given the high level of aggregation at which the alternative household response probabilities are derived. It is in general desirable to know what is the limit of the alternative household estimates based

adjustment.

Regarding the apportionment-based approaches, the following work is outstanding:

- Determine whether obtaining the alternative household estimates is possible for the hard-to-count tiers in practice. If not, a different method of incorporating the Coverage survey information is needed (some were considered);
- Tuning of hard-to-count tiers definition for each of the region, careful collapsing;
- Determine if household structure variable can be defined in such a way that makes it less susceptible to data collection errors.
- Figure out what's going on with within household assisted approach (it looks promising, but something is not as good as it could be.)

As it stands, we recommend to use the direct method with hard-to-count tiers in the 2021 Census unless the further research demonstrate that the within household assisted approach (with apportionment and tiers) can outperform the direct method.

9 Alternative approaches

Other ways of obtaining alternative population estimates for the household bias adjustment may be considered. Here we give a very rough summary of one of such alternatives. There may be some variations of the idea. The idea is to bring in one more data source that can be linked at the address level to the census / coverage survey data and for which population counts by some variable, ideally age-sex, and geography can be produced. It can be an administrative data source. The weighting class approach (Lohr, 2010) can be used in a similar fashion as, say, in Abbott *et al.* (2015) to adjust for the census household non-response (i.e. using census instead of the coverage survey and an alternative source instead of the census as used in the above paper). There are at least two reasons why the weighting classes approach is attractive: (a) it is less sensitive to the overcount error due to the partial cancelation of that error in the ratio and (b) it does not require person level linkage and aggregates within a class across linked addresses. The well-known issue with this approach is the within household non-response in the source being adjusted for the household non-response. So that the relative bias in the weighting class adjusted estimates equals to the proportion of missed individuals within counted households (assuming no other biases present). However, having the coverage survey allows in principle to estimate the census within household non-response and adjust the alternative weighting class estimator for it. Simplistically, the census and alternative data source would be linked at an address level within the coverage survey sampled areas. The census within household undercoverage weights would be estimated by a logistic model fitted into the linked (responding) census and survey households. The weighting class estimator would be applied at the age-sex by survey cluster level (or an aggregation of clusters within a hard-to-count post-stratum) to correct for the census household non-response and within household undercoverage weights applied to

tackle the additional source of error. Next the ratio estimator would be applied using the above cluster level estimates and most likely the census data as the auxiliary to produce alternative estimates by some large area (something like an estimation area) by hard-to-count by age-sex group. Essentially, it is a partial repetition of the 2011 Census coverage estimation with the dual system estimator being replaced by the within-household non-response adjusted weighting class estimator. These alternative estimates alongside the alternative household estimates bases approach then would be used to adjust for the household bias. Some obvious pros and cons. Pros:

- Provide alternative estimates by age-sex;
- No additional person level linkage required compared to certain other alternatives for household or national adjustments;
- Linkage of the census to the alternative source may start before the census to coverage survey linkage starts, so it is a better use of resources;
- The weighting classes method may be a more natural way of reducing the effect of overcount compared to some rule or 'signs of life' based approaches;

Cons:

- A lot more estimators needed to be involved, plus additional assumptions (including independence between census and alternative source), etc.;
- Additional linkage of reasonably high quality is needed, will need clerical resolution;
- There will be some residual overcoverage and heterogeneity bias, both are positive in the case of the weighting classes. Even if there is no household bias in the initial population estimates, the alternative may wrongly suggest that there is;
- Frame dependence between census and alternative source;
- Discrepancies due to different household and address definitions;
- Does not account for the within household bias;
- Requires correctly recorded age-sex info on alternative source;
- Variability of alternative estimates will be substantially higher than variability of initial estimates, would need to think at what level to apply the adjustment;

There is no practical work done on this adjustment approach. It is known that this type of adjustment is also being looked in academia and other statistical institutes (private communication with prof. James Brown).

Bibliography

Abbott, O., Castaldo, A., Račinskij, V., Ross, H., Smith, P.A. & Brown, J. (2015) Developing a weighting-class approach for the 2021 Census. Paper presented at the GSS MAC 29.

Agresti, A. (2002) *Categorical Data Analysis* 2nd. edition. Wiley. New York, USA.

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.

Baffour-Awuah, B. (2009) Estimation of population totals from imperfect census, survey and administrative records. PhD thesis.

Baffour-Awuah, B., Silva, D., Veiga, A. Sexton, C., & Brown, J. (2018) Small area estimation strategy for the 2011 Census in England and Wales. *Statistical Journal of the IAOS*.

Bell, R. B. (1993) Using information from demographic analysis in post-enumeration survey estimation. *Journal of the American Statistical Association*, **88**, 1106-1118.

Brown, J., Abbott, O. & Diamond, I. (2006) Dependence in the 2001 one-number census project. J. R. Statist. Soc. A, **169**, 883–902

Brown, J. and Sexton, C. (2009) Estimates from the census and census coverage survey. GSS Methodology Conference, London, June 2009. ONS.

Brown, J., Abbott, O. & Smith, P. (2013) Design of the 2001 and 2011 census coverage surveys for England and Wales. *Journal of the Royal Statistical Society Series A, 169*, 883-902.

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

Burke, D. and Račinskij, V. (2020) Census coverage survey 2021 sample allocation strategy. *Report to be presented at the Census External Assurance Panel on 24 March, 2020*.

Fienberg, S. (1972) The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables. Biometrika, **59**, 591 – 603.

Office for National Statistics (2012) Household bias adjustment (2011 Census Evaluation Report). Office for National Statistics.

Lohr, S. (2010) Sampling: Design and Analysis. Brooks / Cole, Boston, USA.

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.*

Račinskij, V. (2019) Estimation of the household population in 2021 Census of England and Wales: initial ideas and results. Internal ONS report. Available on request.

Račinskij, V. & Hammond, C. (2019) Overcoverage Estimation Strategy for the 2021 Census of England and Wales. *Report presented at the Census External Assurance Panel on 17 October, 2019.*

Račinskij, V. (2020) Dealing with informative sampling in 2021 Census of England and Wales. Internal ONS report. Available on request.

Wolter, K. M. (1986) *Some Coverage Error Models for Census Data.* Journal of the American Statistical Association, **81**, 338-346.

Appendix (all results for the methods without apportionment

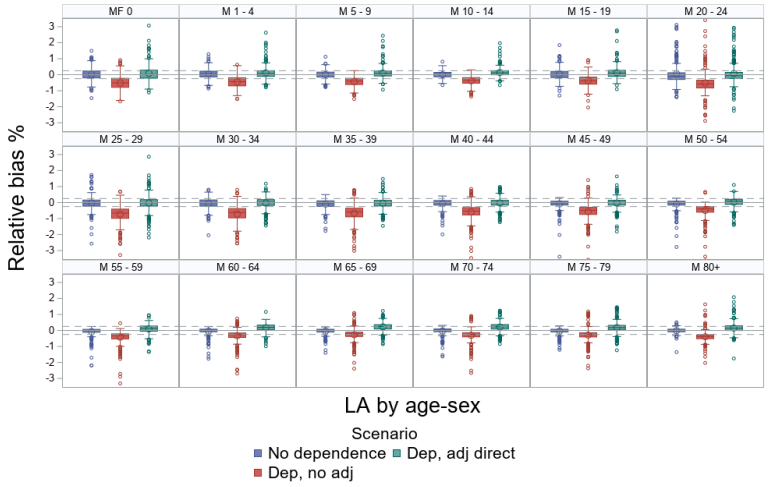

Figure 18: Relative bias, age-sex by local authority totals, males, low dependence

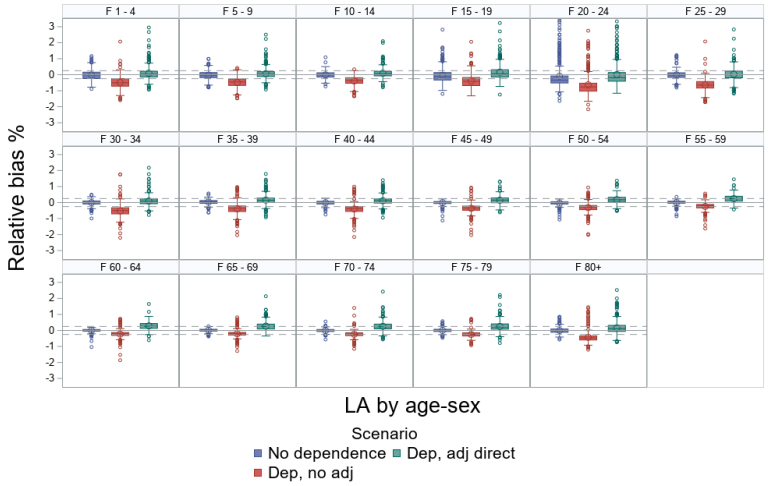

Figure 19: Relative bias, age-sex by local authority totals, females, low dependence

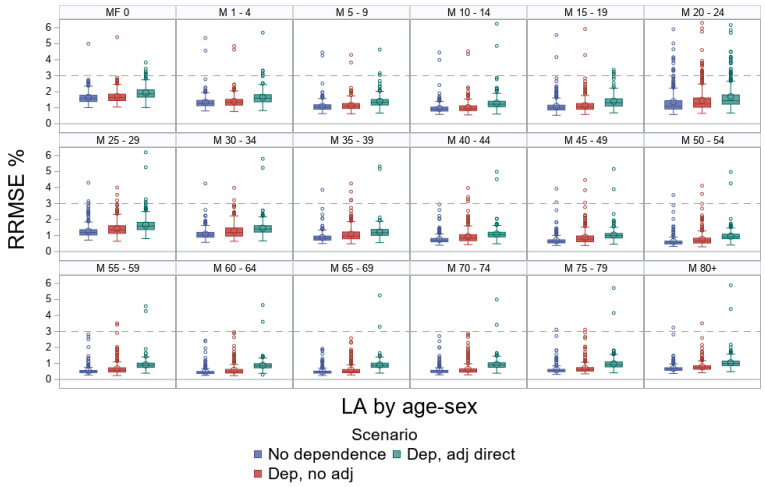

Figure 20: Relative root mean square error, age-sex by local authority totals, males, low dependence

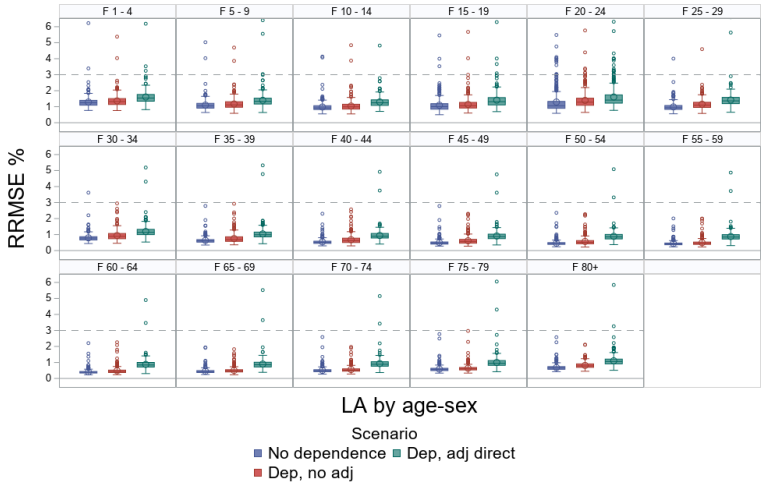

Figure 21: Relative root mean square error, age-sex by local authority totals, females, low dependence

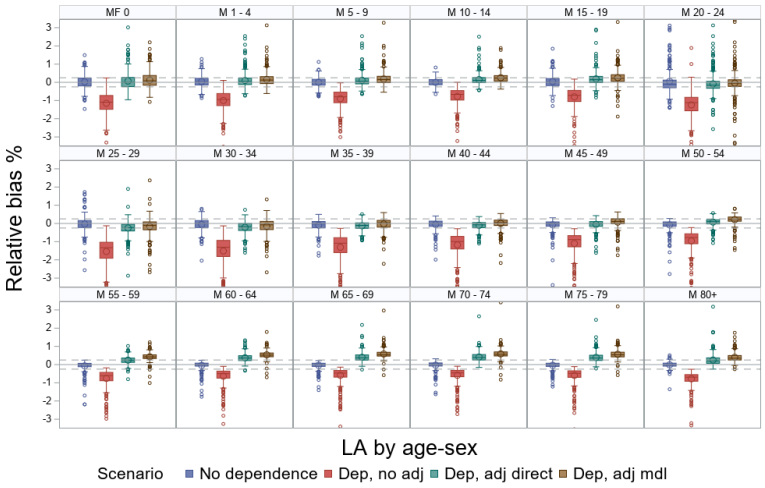

Figure 22: Relative bias, age-sex by local authority totals, males, medium dependence

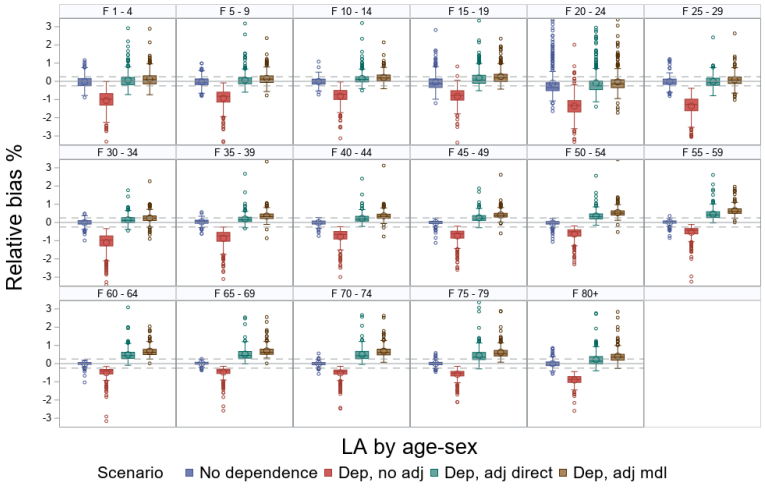

Figure 23: Relative bias, age-sex by local authority totals, females, medium dependence

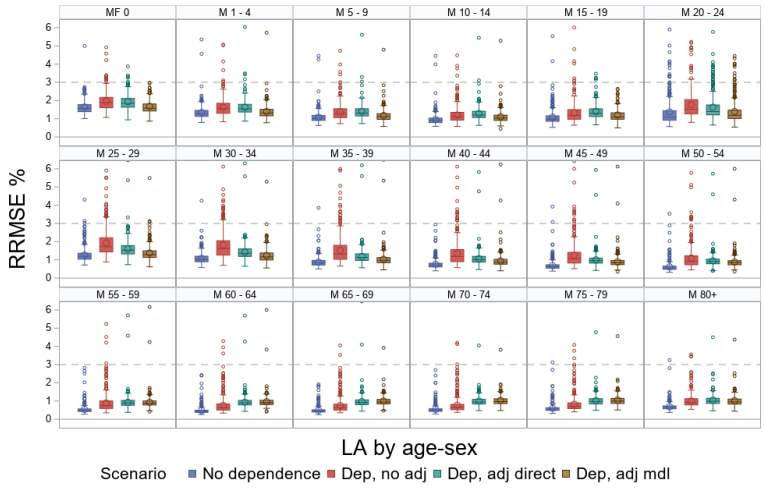

Figure 24: Relative root mean square error, age-sex by local authority totals, males, medium dependence

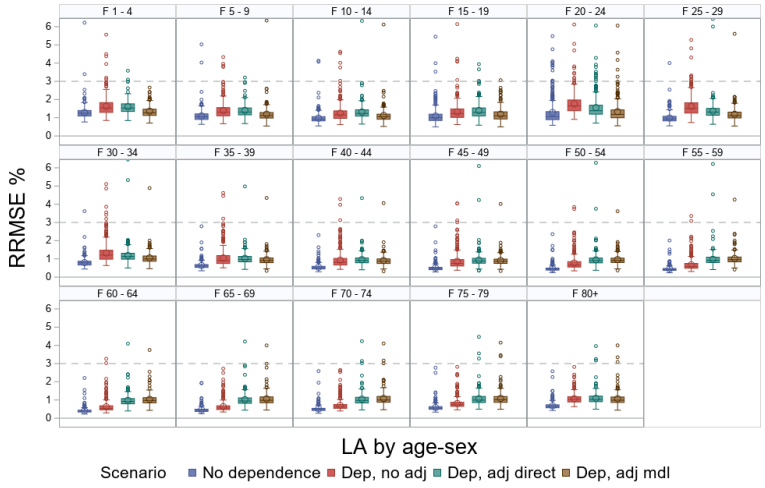

Figure 25: Relative root mean square error, age-sex by local authority totals, females, medium dependence