Dealing with informative sampling in the coverage estimation of the 2021 Census of England & Wales

Viktor Račinskij

Office for National Statistics
Titchfield, UK

February 21, 2020
Draft version 0.5
Disclaimer: all results are preliminary and subject to revision

Executive Summary

Informative (or non-ignorable) sampling occurs when the probability of a population element being selected in a sample correlates with the model's outcome variable. This results in a model that may hold for the sample but does not hold for the population, which leads to incorrect inference. This paper discusses the approaches for dealing with informative sampling in the coverage estimation for the 2021 Census of England and Wales should this type of selection occur.

In the 2011 Census of England and Wales, the Census Coverage Survey was a stratified two-stage cluster design, stratified by local authority and hard-to-count. Coverage estimation was done at estimation area level (aggregate of local authorities batched together based on their similarities and data delivery), which meant the use of optimal allocation was well-justified. In 2011, a way of dealing with informative sampling was proposed by using the design-unbiased model-assisted ratio estimator, however it was deemed less efficient than the model-dependent ratio estimator. The choice in favour of the model-dependent ratio estimator was made in 2011.

For the 2021 Census of England and Wales, the estimation based on generalized linear models for binary responses are proposed. If stratification and the sample allocation approach remain as they were in the 2011 Census Coverage Survey, there will be substantial variability of the sample selection probabilities between the strata as the models are going to be fitted to the entire Census data.

The approaches to be considered are: (1) avoid informative sampling by allocating the sample proportionally; (2) the weighted regression approach, which takes into account the unequal probability sampling; or (3) condition on the variables that are associated with the outcome variable and the sample selection, which will usually involve conditioning on the design variables.

A simulation study is used to consider the outlined approaches. Four estimation approaches (logistic regression, weighted logistic regression, mixed effects logistic regression and mixed effects logistic regression with appended vector of design variables) are considered across two simulation scenarios and two allocation strategies (proportional and optimal allocation).

This paper demonstrates that, if needed, informative sampling selection can be dealt with in the 2021 Census coverage estimation, with the use of the weighted regression approach or conditioning on design variables and their interaction. If optimal allocation (of any type) is chosen, then some more analysis will be done to evaluate the between strata variability.

Does the panel agree that the evidence presented in this paper shows that informative sampling needs to be accounted for in the context of coverage estimation for the 2021 Census of England and Wales? If so, which method would be deemed the most preferred?

1 Introduction

Census coverage estimation aims to produce the census coverage error adjusted population totals at the national and various small domain levels, say, a local authority or a local authority by an age-sex group. These totals are often achieved by combination of the survey sampling, capture-recapture, predictive modelling and small-area estimation methods, see for example Brown *et al.* (2019) and Baffour *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) and Račinskij & Hammond (2019).

Similarly to the coverage estimation of the 2001 and 2011 Censuses, there will be a large post-enumeration survey, known as the Census coverage survey, at the core of the coverage estimation of the 2021 Census of England and Wales. It can be argued, that the coverage error adjusted totals are technically survey-based estimates with the census data being the high quality auxiliary data. Since the census coverage survey in general has a complex sampling design, the data analysis may require taking the design related features into account. This report discusses one of such features usually referred to as *informative* (Chambers & Skinner, 2003; Pfeffermann, 2011) or *non-ignorable* sampling (Valliant *et al.*, 2000; Gelman, 2007).

Simplistically, a sampling process is informative whenever the sample selection probability of a population element correlates with the model's outcome variable. As a result, the model may hold for the sample, but not for the population leading to incorrect inference. Informative sampling and the ways to deal with it is a vast area in statistics. However, in this report we limit the discussion to the essential minimum needed in the context of the coverage estimation. Note also that we do not discuss many other topics in the analysis of the complex survey data such as reflecting the cluster design, for instance. To make our analysis slightly more straightforward, we also ignore the survey non-response in our discussion.

2 Informative sampling: an outline

Arguably, very few surveys employ the simple random selection without replacement scheme in practice. Specifically, to increase the efficiency of estimation, stratification is often used. Stratification need not in general lead to an unequal probability sampling, but frequently the sample selection probabilities are indeed unequal. Weighting in the design-based methods explicitly accounts for the complex design. However, when the model-based approaches are used, it is quite common to assume that the data were generated directly from the population of interest (Chambers & Skinner, 2003; Gelman, 2007).

When the sample selection probabilities are unequal at some stages of sampling and the outcome variable is correlated to the selection probabilities, the sample density and the population density need not be equal. In such situation the sampling is called informative (or non-ignorable) and the subsequent inference may not be applicable to the population.

To put it slightly more formally, let y_i be an outcome variable for an i^{th} element, $i = 1, \ldots, N$. Let $\boldsymbol{x}_i$ be a vector of covariates and I_i be the sample membership indicator, f_s and f_U be the sample and the population densities, respectively. Then, as shown in Pfeffermann (2011), we have the following relationship

$$f_s(y_i \mid \boldsymbol{x}_i) = f_U(y_i \mid \boldsymbol{x}_i, I_i = 1) = \frac{P(I_i = 1 \mid y_i, \boldsymbol{x}_i) f_U(y_i \mid \boldsymbol{x}_i)}{P(I_i = 1 \mid \boldsymbol{x}_i)}. \tag{1}$$

It is easy to see that the two densities are equal if the following condition is satisfied:

$$P(I_i = 1 \mid y_i, \boldsymbol{x}_i) = P(I_i = 1 \mid \boldsymbol{x}_i). \tag{2}$$

In other words, conditional on the covariates the response and the selection must be independent in order for the sampling selection be non-informative or ignorable (Zimmermann, 2018). This means that all the variables that are simultaneously related to the sampling and outcome of interest should be included (Gelman, 2007). This echoes the notion that the bias due to informative sampling is incurred when both the sampling fractions and the values of the response variables are different within an estimation stratum in the finite population predictive approach (Valliant *et al.*, 2000), see Brown *et al.* (2019).

Note that there are situations when even strongly informative sampling can be ignored. For instance, when the logistic regression is used in the case control studies and only the slope parameters are of interest (Agresti, 2002). Unfortunately for us, the census coverage estimation relies on the descriptive analysis where the intercept parameter is essential for the reliable results. Moreover, it has been demonstrated in Pfeffermann and Sverchkov (2003) that if sampling selection also depends on $\boldsymbol{x}_i$, the slope parameters are also affected by the informative sampling. Our research shows that it is also the case in the coverage estimation.

3 Background on the design of the Census coverage survey

Similarly to the 2011 Coverage survey, the 2021 Census coverage survey is expected to have the size of 335,000 households across England and Wales. The proposed design is similar to the design of the previous Coverage survey (Brown *et al.*, 2011): stratified two-stage cluster. The sample is stratified by local authority by hard-to count. Within each stratum a simple random sample without replacement of output areas is going to be selected. The first stage sampling fractions are expected to vary between the strata. Within each selected output area a simple random sample without replacement of the postcodes will be selected using a certain fixed sampling fraction. At the first stage of sampling, the x-optimal allocation (Särndal *et al.*, 1992) with minimum and maximum sample size constraints was used in 2011. Sample allocation strategy is yet to be decided for the 2021 Coverage survey and it partly depends on the result of this work.

Note that the stratification involves the hard-to-count index which essentially reflects the level of the census non-response in the previous census. In the coverage estimation, we use the weights that are reciprocals of estimated census coverage probabilities (Alho, 1990; US Census Bureau, 2008; Račinskij, 2018) . It is therefore obvious how an association between the outcome variable and the sample selection process may occur. If the sample allocation process is disproportional, we are having the informative selection unless we carefully condition on all the variables related to the census response and the sample selection.

4 Coverage estimation in the 2011 Census of England and Wales and informative sampling

In the coverage estimation of 2011 Census a combination of the dual system, ratio and synthetic estimators was used, see Brown *et al.* (2019) for details. Here, we are mainly interested in the way the 2011 estimation method is related to the topic of the report.

Since the data were delivered in batches in 2011, the optimal allocation described above was well justified. The dual system estimator was applied at the cluster of postcode by an estimation variable, by a hard-to-count stratum by a local authority to provide protection against the heterogeneous inclusion probabilities:

$$\hat{t}_{vLhc} = \frac{(X_{vLhc} + 1)(Y_{vLhc} + 1)}{(M_{vLhc} + 1)} - 1, \tag{3}$$

where X_{vLhc} is the census count for an estimation variable v, a local authority L, a hard-to-count group h and a sample cluster c; Y_{vLhc} is the corresponding survey count and M_{vLhc} is the corresponding census to survey match count. Local authorities were batched together into groups known as estimation areas based on their similarities and the data delivery timetable. The model-dependent ratio estimator was applied at the variable by estimation area by hard-to-count level:

$$\hat{T}^{(md)}_{vEh} = \frac{\sum_{L\in E}\sum_{c\in s_{Lh}} \hat{t}_{vLhc}}{\sum_{L\in E}\sum_{c\in s_{Lh}} X_{vLhc}} \sum_{L\in E} X_{vLh}, \tag{4}$$

where E is estimation area, s_{Lh} is census coverage survey sampled areas in Lh.

For the local authorities large enough to be an estimation area on their own, there was no risk of estimation being affected by the informative selection, since the sampling and estimation strata were coterminous in this case. When an estimation area was made of several local authorities, there was a potential risk of sampling being informative. In this case, if sampling fractions and response propensities varied simultaneously between the local authorities within the estimation area, the bias would had been incurred (Brown *et al.*, 2019). The way around it would be using the model-assisted ratio estimator:

$$\hat{T}_{vEh}^{(ma)} = \frac{\sum_{L\in E}\sum_{c\in s_{Lh}} w_{Lh}\hat{t}_{vLhc}}{\sum_{L\in E}\sum_{c\in s_{Lh}} w_{Lh}X_{vLhc}} \sum_{L\in E} X_{vLh}, \tag{5}$$

where w_{Lh} is the design weight for the local authority by hard-to-count. Therefore, (5) would be design-unbiased, but less efficient than (4), whereas latter is not necessarily design-unbiased. The choice in favour of (4) was made in 2011.

5 Coverage estimation in the 2021 Census of England & Wales and informative sampling

In the coverage estimation of 2021 Census of England & Wales the estimation based on the generalized linear models for a binary response are proposed (Račinskij, 2018; Račinskij & Hammond, 2019). It is anticipated that these models will be fitted into the linked Coverage survey to census data for the entire England & Wales.

If the stratification and sample allocation approach remains as it was in the 2011 Coverage survey, there will be substantial variability of the sample selection probabilities between the strata. There is a risk of sampling being informative as the outcome variable (whether a survey responding individual / household responded in census, $y_i = \{0, 1\}$) is associated with the stratification by hard-to-count. If not accounted for, it may incur the bias.

We have briefly discussed why the stratification / allocation strategy was justified in the 2011 coverage estimation. Would the same allocation strategy be well justified with the estimation method proposed for the 2021 Census? On the one hand, the fact that the modelling approach will borrow the strengths across the areas reduces the need for allocating the sample by local authority and hard-to-count. In addition, operationally the 2001 and 2011 Censuses were a lot more similar, whereas the 2021 Census operationally will differ from the two. Therefore, using the response patterns at the local authority by hard-to-count from the previous census may give less gains than it gave in 2011. Of course, broadly, the response will vary by the hard-to-count and census response in 2021 will be correlated with the variable. On the other hand, allocating sample proportionally will sort the informativeness of the sampling procedure and will be justifiable from the pure modelling perspective. Nevertheless, there is a number of points to consider. First of all, there is differential survey non-response by hard-to-count which will result in disproportion in the realised sample. Hence, some disproportional allocation may mitigate for it. Also, despite borrowing the strength, we may not be able to achieve the comparable level of precision for some of interactions of hard-to-count with other variables in the case of proportional allocation, for instance. So in principle, we may be better off with some compromise allocation, but this needs more work to be done (Burke and Račinskij, 2020).

6 Simulation study

We conducted a simulation study to explore the effect of the informative sample selection on the coverage error corrected population size estimation. These simulations are a further development of the Brown and Sexton (2009) and those used in Račinskij (2018) and Račinskij (2019), but have some slight modifications needed for the problem under consideration. In these simulations the census coverage and the census data are generated from the models fitted to the 2011 coverage survey to census linked 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). In addition, the model contains a random intercept term. Unlike the simulations referenced above, the random intercept in this study is not at the local authority level, but at the sample stratum level, that is at, a local authority by hard-to-count level. The Coverage survey model is similar, but categorical age-sex effect and a random intercept at the local authority level is used. Data generated so that the perfect linkage, closed population and independence assumptions are satisfied. There is no overcoverage in this study.

Two census scenarios are used. The first one, referred to as the *base* scenario, is directly generated from the model fitted to the real 2011 data. The second one, referred to as the *excess* scenario, uses the original fixed effects of the above models (like the *base* scenario), but modifies the random local authority by the hard-to-count residuals. Modification proceeds is as follows: all the original residuals are sorted by local authority and hard-to-count and then systematic random sampling selects every second residual, keeps its original sign but raises its absolute value to the power of 0.3 (an arbitrary figure picked to produce large enough variability between the strata); all non-sampled residuals remain unchanged. In this way we generate two populations that are identical in terms of the underlying fixed effects, but in the second case we have a lot more variable response patterns between the sampling strata. In the *excess* scenario there is substantial variability in hard-to-counts within a local authority as well as between local-authorities within a hard-to-count. Note, that the response rates in the *excess* decrease in the quarter of cases, remains unchanged in the half of cases and increases in the quarter of cases compared to the *base* scenario. The overall response rate is around 0.94 in both scenarios (slightly lower in the excess compared to the base). It must be noted, that variation introduced in the *excess* scenario is quite large and may not be realistic, but it is helpful to highlight possible issues with informative sampling.

The sampling design in this study is as in the 2011 Coverage survey, we use two allocation strategy: proportional and optimal. In the case of the optimal allocation it is truly optimal as we use the exact coverage probabilities (including the contribution of the random effects) in the allocation process. Due to technical issues, the number of simulation scenarios had to be limited to 128.

7 Approaches

It is not always straightforward to achieve (2) in practice. There are numerous approaches of dealing with the complex survey data which may vary depending on the context (Chambers & Skinner, 2003). In this report we will focus on the most relevant in the situation of the census coverage estimation. Also, given that the coverage estimation process is already reasonably complicated when the informative sampling is ignored, we tend to focus on the methods that are easier to implement using existing software.

We will consider the following ways of dealing with the above issue:

1. Avoid the informative sampling. Since the Census coverage survey is specifically designed for the coverage estimation, we have control over its design. Certain sampling schemes result in the non-informative sampling. The most notable, of course, is the simple random sampling without replacements, which is not really feasible in the case of the Coverage survey. However, broadly keeping the design of the 2021 Coverage survey similar to the design of the 2011 Coverage survey, but allocating the sample proportionally, will result in approximately non-informative selection.

2. Use the weighted regression approach. The weighted regression takes into account the unequal probability sampling when solving the estimating equations and its outcomes are design consistent for the population (or census) parameter (Skinner, 2003).

3. Condition on the variables that are associated with the outcome variable and the sampling selection. This will usually involve some form of conditioning on the variables used in the sample selection, $\boldsymbol{z}_i$, commonly referred to as the design variables. Note, it is argued, that not only $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ should be included in the model, but also their interactions (Pfeffermann, 2011). Clearly, it may not be always feasible to do so.

Four estimation approaches are considered across the two simulation scenarios and two allocation strategies. All of them involve some form of the logistic regression. Note, that the logistic regression is used for convenience and speed, but there are no reasons why a different link function cannot be used. In all four cases the reciprocal of the estimated census response probability is used as the non-response weight.

The first approach is based on the logistic regression. It includes categorical age-sex group, activity last week, accommodation type, address one year ago, household relation, marital status, person ethnicity, region, self-contained accommodation, tenure and the following second order interactions: ethnicity with tenure, ethnicity with region and activity last week with region. We use the same vector of the fixed effects $\mathbf{x_i}$ in all approaches. There is also a vector $\mathbf{z_i}$ of rudimentary design variable that contains a hard-to-count variable with the standard 5 levels. Hence, design variable does not reflect the crossing of the local authority and hard-to-count in this case. The total for a variable v in a local authority L is estimated as

$$\hat{T}_{vL}^{LR} = \sum_{r \in vL} \hat{\pi}_r^{-1} = \sum_{r \in vL} \left[\frac{1}{1 + \left(- \left[\mathbf{x}_r^T \hat{\beta} + \mathbf{z}_r^T \hat{\kappa} \right] \right)} \right]^{-1}, \tag{6}$$

where $r \in vL$ means that an element r belongs to the study variable and is located in an area L; $\hat{\pi}_r$ is the estimated census response probability.

The second method is the weighted logistic regression, it uses exactly the same predictors as (6) and is given by

$$\hat{T}_{vL}^{WLR} = \sum_{r \in vL} \hat{\rho}_r^{-1} = \sum_{r \in vL} \left[\frac{1}{1 + \left(- \left[\mathbf{x}_r^T \hat{\beta}^{(w)} + \mathbf{z}_r^T \hat{\kappa}^{(w)} \right] \right)} \right]^{-1}, \tag{7}$$

where $\hat{\rho}_r$ is the estimated census response probability. We use various superscripts over the estimated coefficients (like (w) in the above case) to highlight the fact that these estimated coefficients will in general vary between different approaches. This approach was run for the optimal allocation only.

The third method is the mixed effects logistic regression, with the same predictors as in (6) and (7), but also having a random intercept at the local authority level

$$\hat{T}_{vL}^{MELR} = \sum_{r \in vL} \hat{\tau}_r^{-1} = \sum_{r \in vL} \left[\frac{1}{1 + \left(- \left[\mathbf{x}_r^T \hat{\beta}^{(m)} + \mathbf{z}_r^T \hat{\kappa}^{(m)} + \hat{u}_L \right] \right)} \right]^{-1}. \tag{8}$$

The final approach is as (8), however, we have appended the vector of the design variables to $\mathbf{z_i^*}$ which now contains the hard-to-count index as before plus the two continuous variables used in sample allocation and an interaction of these two variables: stratum size within a local authority by hard-to-count and the variance of the imputation rates in the 2011 Census within a local authority by hard-to-count. Note, that it may not be exactly the way one would be conditioning in the real application, but it was the most straightforward way in terms of implementing it in the study. The estimator is given as

$$\hat{T}_{vL}^{MELRz} = \sum_{r \in vL} \hat{\xi}_r^{-1} = \sum_{r \in vL} \left[\frac{1}{1 + \left(- \left[\mathbf{x}_r^T \hat{\beta}^{(z)} + \mathbf{z}_r^{*T} \hat{\kappa}^{(z)} + \hat{u}_L^{(z)} \right] \right)} \right]^{-1}. \tag{9}$$

It would be very natural to consider (6) with $\mathbf{z_i^*}$ instead of $\mathbf{z_i}$. Unfortunately, due to the time constraints we did not look at this option at this time. Nevertheless, it is easy to guess its performance based on the performance of (9). Note also that (9) was run for the excess scenario for the same reason.

In order to avoid biases mixing up, the Coverage survey has 100% response rate in our simulations. This removes any heterogeneity bias. Then essentially we are left with the bias due to model misspecification / syntheticity and the design bias.

Note also, that none of the above approaches is tuned to give the ultimate performance, therefore we have more bias due to model misspecification than one would anticipate in reality.

Note also, that none of the above approaches is tuned to give the ultimate performance, therefore we have more bias due to model misspecification than one would anticipate in reality.

8 Results

We look at the results now. We start with the high level totals since the bias due to informative sampling will be mostly noticeable at such level. The relative bias at the regional level is presented in Figure 1 for the base scenario, and in Figure 2 for the excess scenario. For the base scenario, the weighted regression based estimator under the optimal allocation and the mixed effects based approach with the proportional allocation perform really well. The performance of the wighted regression is exactly as expected in this situation because it ensures a good population level model. The logistic regression based approach and the mixed effects regression based approach with the optimal allocation tend to overestimate the population size slightly. This overestimation is a consequence of the informative sampling under which disproportionately more 'harder' respondents are sampled leading to the estimated probabilities of the census non-response being lower in the sample than they are in the population. Hence, the non-response weights are larger in the sample than they should be in the population.

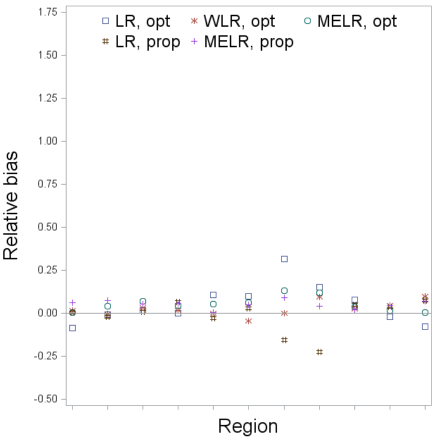

Figure 1: Region totals, relative bias %, base scenario

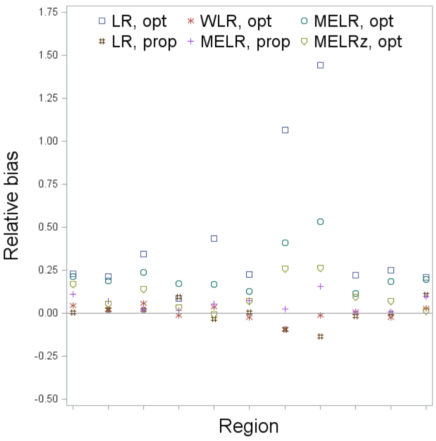

Figure 2: Region totals, relative bias %, excess scenario

In the excess scenario the trend is similar, but the bias is of larger magnitude. This is due to sharp increase in correlation between the sample selection probability and response. The mixed effects logistic regression based approach with appended vector of design variables (9) performs well, but not as good as the weighted regression.

In terms of the relative root mean square error (Figure 3 and Figure 4, base / excess scenarios, respectively) for the regional totals, the mixed effects based approaches and the weighted regression approach perform a bit better than the rest in majority of the cases in the base scenario. In the excess scenario, the performance of (7) and (9) is obviously better than the performance of the remaining estimators. It is interesting to see, that the weighted regression performs that well given that the design weights are lot more variable in this case.

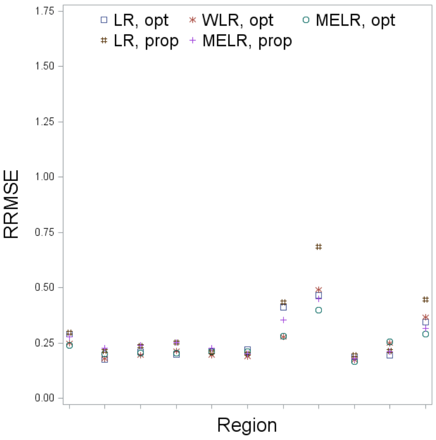

Figure 3: Region totals, relative root mean square error %, base scenario

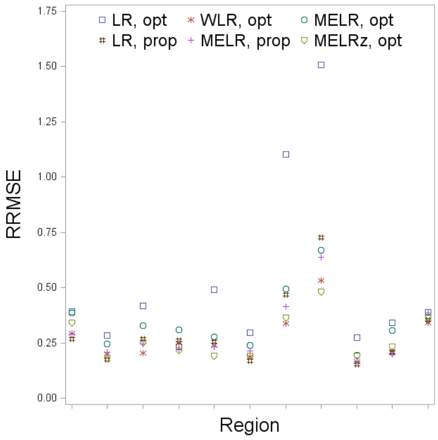

Figure 4: Region totals, relative root mean square error %, excess scenario

Looking at the relative bias of the age-sex totals in the Figure 5 for the base scenario and Figure 6 for the excess scenario, we see the similar patterns seen in the regional estimates. In both the base and excess scenarios the weighted regression with the optimal allocation and approaches with the proportional allocation outperform those with the optimal allocation. Again, in the excess scenario the impact of the informative sampling is quite prominent. We see that the conditioning on the appended design variables approach (9) works reasonably well.

In terms of the relative root mean square error (Figure 7 and Figure 8, base / excess scenarios, respectively) for the age-sex group totals, it becomes difficult to single out the better performing approach in the base scenario. However, in the case of the excess scenario, (7) and (9) perform better than the rest.

Looking at the relative bias for the local authority totals, Figure 9, we see how the logistic regression approach slightly overestimates in each local authority compared to the weighted logistic regression based approach. Of course, we do not really expect to see many easy noticeable differences between approaches at the local authority level. The same data in the form of box-plots are presented in Figure 10 where the slight shift for the mean of the distributions can be seen with the smallest shift in the weighted regression based approach.

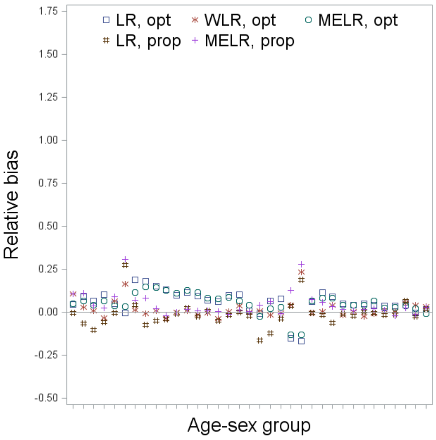

Figure 5: Age-sex totals, relative bias %, base scenario

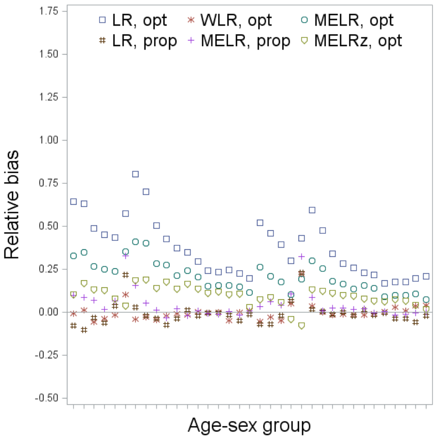

Figure 6: Age-sex totals, relative bias %, excess scenario

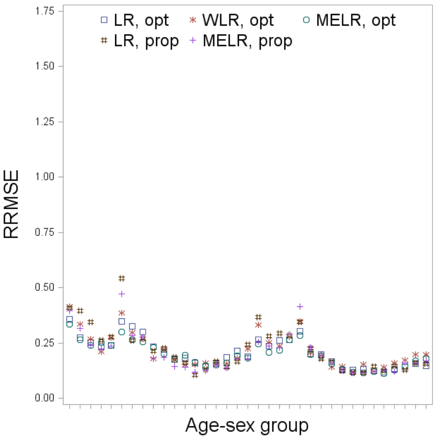

Figure 7: Age-sex totals, relative root mean square error %, base scenario

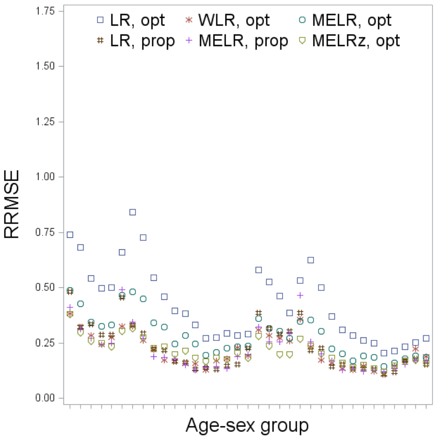

Figure 8: Age-sex totals, relative root mean square error %, excess scenario

The relative root mean square errors for the local authority level estimates are presented in Figure 11. The logistic regression and the weighted logistic regression have better performance on this metric for the local authority estimation, with the weighted regression arguably being slightly better. We see that the mixed effects regression based approaches quite often have a small bias but larger variances at the local authority level. As the level of aggregation increases, the bias starts to dominate and the mixed effects based approaches have better performance at the higher levels. Clearly, better conditioning should increase the performance of the fixed effects only models in cost of some variance. Note also, that there is relatively little difference in the performance of the mixed effects models in the base and the excess scenarios. Note also the effect of the optimal allocation on the relative root mean square error for the mixed effects based approaches.

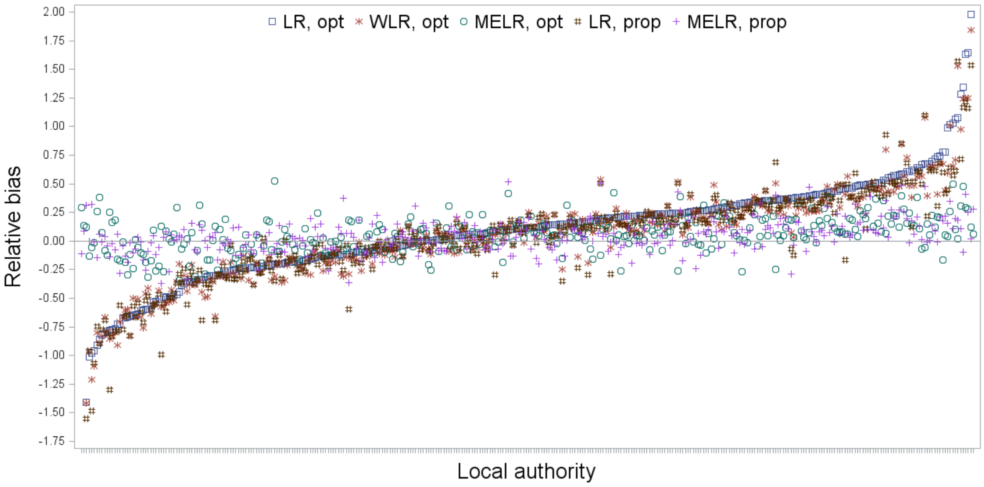

Figure 9: Relative bias, local autority totals, base scenario (note that some values where excluded, compare with Figure 10)

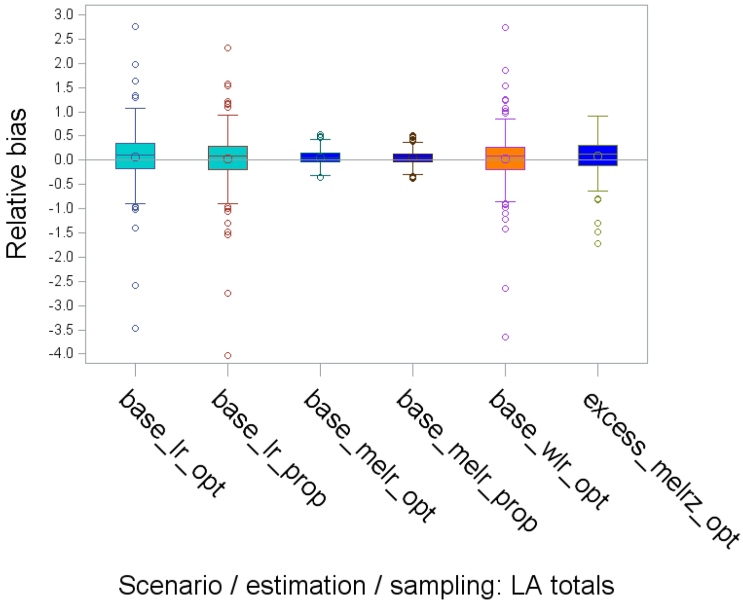

Figure 10: Relative bias, local autority totals, base scenario and MELRz in the case of excess scenario

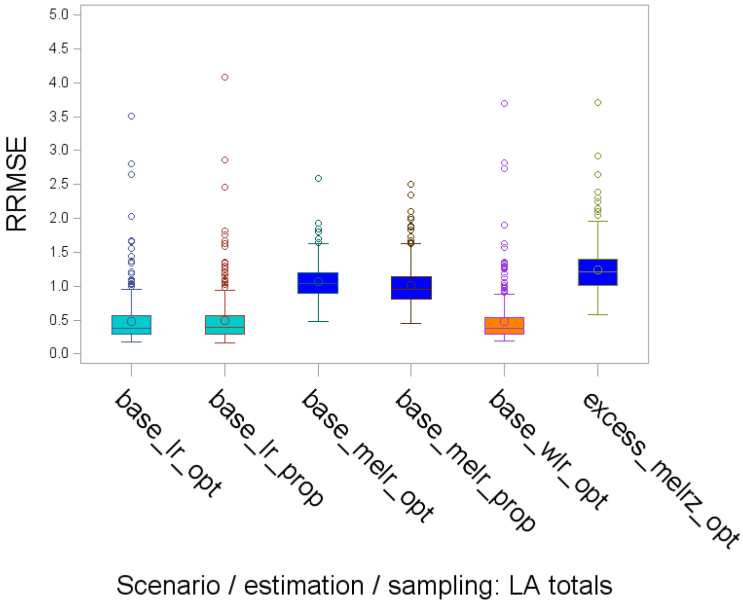

Figure 11: RRMSE, local autority totals, base scenario and MELRz in the case of excess scenario

9 Conclusions, future work and recommendations

It has been demonstrated that, if needed, we would be able to deal with the informative sample selection in the 2021 Census coverage estimation. Both the weighted regression and the approach of conditioning on the design variables and their interactions showed acceptable performance. Note that this may be only true for the sampling design discussed in the section 3. There is no guarantee that under more complex sampling design or allocation the methods considered in this work would produce reliable results.

If an optimal allocation of any type is chosen, there will be a need for some analysis in order to carefully evaluate the amount of the between strata variability. However, the expectation is that the between strata variability will be close to the one observed in the base scenario.

The remaining work is to assess the performance of the estimation approaches discussed in this paper with the compromise sample allocation (optimal at the hard-to-count level, proportional within each hard-to-count for the local authorities).

Our recommendation for the coverage estimation is to build a fixed or mixed effects model with all design effects first and then to perform diagnostic tests (Chambers & Skinner, 2003) for the informativeness of the sampling. If a substantial effect is detected, we recommend switching to the weighted regression model with the same sets of covariates as in the initial model (possibly dropping some of the design effects).

Bibliography

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, 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.*

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. and 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.*

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

Chambers, R. L. and Skinner, C.J. (2003, Eds.) *Analysis of Survey Data*, Wiley, New York, USA.

Gelman, A. (2007) Struggles with Survey Weighting and Regression Modeling *Statistical Science, 27(02)*, 153-164.

Pfeffermann, D. and Sverchkov, Yu. (2003) Fitting Generalized Linear Models under Informative Sampling. In Chambers, R. L. and Skinner, C.J. (2003, Eds.) *Analysis of Survey Data*, Wiley, New York, USA.

Pfeffermann, D. (2011) Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology, 37(2)*, 115-136.

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.*

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling.* Springer-Verlag, New York, USA.

Skinner, C. (2003) Introduction to part B. In Chambers, R. L. and Skinner, C.J. (2003, Eds.) *Analysis of Survey Data*, Wiley, New York, USA.

US Census Bureau (2008) 2010 Census Coverage Measurement Estimation Methodology. US Census Bureau, Washington, D.C.

US Census Bureau (2012) 2010 Census Coverage Measurement Estimation Report: Aspects of Modeling. US Census Bureau, Washington, D.C.

Valliant, R., Dorfman, A. and Royall, R. (2000) *Finite Population Sampling and Inference: a prediction approach.* Wiley, New York, USA.

Zimmermann, T. (2018) The interplay between sampling design and statistical modelling in small area estimation. PhD thesis.