Weighting class adjustment and population size estimation for Census 2021 – a simulation study

James Dawber, University of Southampton, UK.

Key Messages of Paper

Purpose

- This paper explores the properties of a weighting class estimator in the context of a Census, administrative data and a survey. The paper shows how various scenarios impact the estimators in terms of bias and variance. It is compared to a standard DSE approach, so that tradeoffs between the two methods can be clearly shown.

Recommendation

- The paper shows that the weighting class estimator, while robust against over-coverage, makes very strong assumptions about within-household non-response which are not ignorable and result in bias. In addition, its variance is in general higher than that from a DSE approach. Essentially, it is a potential option only in specific scenarios where the biases in a DSE cannot be estimated and adjusted for.

Key Asks of MARP

- The panel is asked to:
    - Note the results, in the context of the potential use of these methods as part of the census playbook, and the potential use in the future population statistics system.
    - Provide feedback on the study.

0 Executive summary

- The advantages of using weighting class adjustment (WCA) on a Census using an admin list compared to dual system estimation (DSE) is that it is robust to over-coverage and list dependence.
- The disadvantage of WCA compared to DSE is that it is sensitive to Census within-household under-coverage.
- The WCA underestimates population sizes primarily due to within-household under-coverage in the Census and this is determined by both individual and household level factors, beyond just age-sex characteristics. Even with an error-free and complete admin list, WCA will get negatively biased estimates due to household-level linking and Census within-household under-coverage.
- The primary disadvantage with DSE is that it is sensitive to over-coverage in the admin list, leading to over-estimation of population sizes. A secondary disadvantage is that list dependence between the Census and admin list contributes to under-estimation but is unlikely to have a biasing effect as significant as over-coverage.
- The simulation study shows that if within-household under-coverage is greater than the expected over-coverage, then the WCA estimate will be more biased than DSE. The study also suggests that it is likely that the true population size will be between the WCA estimate and the DSE, if under-coverage is similar to the 2011 Census and over-coverage is non-trivial.

1 Introduction

Non-response to the Census causes under-coverage, complicating the accurate estimation of population sizes. In previous censuses, a coverage survey has been used to measure and adjust for this under-coverage. In recent years, admin-based population estimates (ABPE), utilising four administrative data sources, have been developed. The quality of the ABPE have improved over time, yet are still not accurate enough to be trusted outright. However, they may be accurate

enough to provide the means to provide accurate estimates when used in conjunction with the Census. One such method is to use weighting class adjustment (WCA) to reweight certain demographic classes proportional to the under-coverage in the Census, leading to accurate population size estimates without the need for a coverage survey. The purpose of this study is to assess whether WCA may be a viable approach to adopt in the 2021 Census. This assessment is made using a comprehensive simulation study.

The aims of this simulation study are as follows:

1. To assess whether it is likely that WCA can accurately estimate population size alongside a Census.
2. To assess under what conditions WCA would be suitable for use with a Census.
3. Assess how well WCA performs in realistic but non-ideal scenarios where:
    a. Non-response in the Census and the admin list are heterogeneous within and between classes.
    b. Non-response in the Census is not independent of the admin list.
    c. The list has over-coverage.
4. Compare WCA to an alternative approach using dual system estimation (DSE).

To assess whether WCA can be useful in practice, it is important to ensure that the simulation is as close to reality as possible. To do so, we utilise ONS data from the 2011 Census and recent ABPE publications to help construct realistic populations and scenarios. This will mean that many complexities will be introduced that will likely bias the WCA estimates to some extent. By observing this bias in different simulation scenarios, we can assess the likelihood of WCA being a potentially useful method to use with the 2021 Census.

The report is structured as follows. In the next section the methods that were used in the simulation are outlined. Next, the results of the simulation are presented, and finally, the conclusions are made in the final section.

2 Simulation methods

The simulation study requires two general steps to ensure it is comparable to reality. The first step is generating a synthetic population similar to the real population, including the classes for WCA. The second is simulating respondents from the synthetic population such that they are representative of a Census and admin list. The WCA estimates can then be calculated using the synthetic Census and admin list, and these estimates compared to the known synthetic population sizes. The classes used for the simulation were age-sex classes for five-year age groups up to 79 years, and 80 or over. We also consider a new-born age-sex class that includes both sexes aged less than one year old. These two steps each have complexities which make it difficult to mirror reality, hence the simulation must rely on simplifying assumptions. In developing the simulation, we aim to minimise these assumptions as much as possible, while maximising the use of available data to mimic reality as closely as possible.

Due to accurate ONS population data, generating the synthetic population is relatively straightforward compared to generating the Census and admin lists. There are many mechanisms to consider in order to accurately do this, including:

1. Generating under-coverage at both the individual and household level.
2. Generating under-coverage which varies across different individual and household groups, e.g., age, sex, household tenure and household size.

3. Generating under-coverage for both the Census and admin list such that it incorporates some degree of dependence between them.
4. Generating over-coverage in the admin list.

For the purposes of this research, it is assumed that census over-coverage is negligible. Details of how the synthetic population was generated, and how under- and over-coverage were generated in the two lists are described below in the remainder of this section. Simulation parameters are also introduced and described that will be varied in the simulation study.

2.1 Synthetic population

A synthetic population with N individuals was created, with the proportions of the population in the respective 35 age-sex classes based on the 2011 Census. Further variables were considered that were based on 2011 Census counts for England and Wales across household tenure and household size. These variables were used because they have also been shown to be associated with Census response rates. The synthetic population of individuals was randomly selected using these 2011 counts as sampling weights.

Each individual then needed to be assigned to a household identifier. To do this, approximately $1/k$ adults from each household tenure and household size k were randomly assigned to a household, and then the remaining individuals, both children and adults, were randomly added to these households ensuring the household sizes remained at k individuals. This was done to ensure that individuals were randomly assigned to households based on their pre-established household size attribute, e.g., two individuals in a household size of two could be paired together. Individuals were not assigned to households based on their age; hence the synthetic population will have more within-household variation than in reality. For example, households with two adults generally will be similarly aged whereas this will not be the case in the simulation. This simple approach is sufficient to explore precision of the estimators in this study. Finally, each individual was randomly assigned a sex with equal probability and a hard-to-count index from 1 to 5, with probability weighting (0.4, 0.4, 0.1, 0.08, 0.02). The higher the value the harder the individual was to count. Since sex and the hard-to-count index were added independently, they have no association with the other variables in the synthetic population. In total, the synthetic population has the following variables:

- Sex
- Age
- Household tenure
- Household size
- Household identifier
- Hard-to-count index

These variables each will have different effects on response probabilities.

2.2 Response probabilities

With the synthetic population created the response probabilities for both the Census and the administrative list can now be calculated. The response probability is equivalent to the complement of the non-response probability. We use the term ‘response probabilities’ for the admin list as well as the Census for consistency, despite admin lists technically not having respondents. For example, if an individual has an admin list response probability of 0.9, there will be a 90% chance that individual will be randomly selected to be in the admin list.

To generate the response probabilities for the Census we use a two-stage approach which first selects responding households, and secondly selects responding individuals within these households. Hence, we need response probabilities at the household level and at the individual level. Household response rates for the 2011 Census varied due to many factors, but for this simulation we focus on the household size and tenure. To utilise this variation in the simulation, we used the 2011 household response rates published by the ONS. Since this source provides just the marginal response rates of household sizes and tenures, we infer the joint response rates by assuming independence between the two variables. For example, if the response rate for households with two people is p_{hhd2} and the response rate for 'social rented' households is p_{SR} then the response probability for a 'social rented' household with two people in it was assumed to be $(p_{hhd2}p_{SR})/p_{Total}$ where $p_{Total} = 0.958$ is the total household response rate. So, for each household h a response probability (p_h) was calculated based on these rates. Responding households were simulated using a *Bernoulli*(p_h) distribution.

To get the response probabilities of individuals within a responding household, an existing regression model was used which utilised individual and household level attributes. Pre-existing logistic regression parameters ($\boldsymbol{\beta}$) were supplied by the ONS that used covariates ($\boldsymbol{X}$) age-sex class, household tenure, household size, and a hard-to-count index as predictors of response probabilities for the 2011 Census. The model-derived individual Census response probabilities ($p_i^{(c)}$) for each individual i were then calculated using these covariates and the provided parameter vector:

$$p_i^{(c)} = logit^{-1}(\boldsymbol{X_i\beta}).$$

Individuals in households that had already been not selected in the Census (with probability $1 - p_h$) had $p_i^{(c)}$ set to zero.

The response probabilities for the admin list will naturally differ from the Census. This is because the capture mechanism is very different. The admin list combines several different administrative records, which makes it difficult to estimate the response probabilities like the Census. To simulate the response probabilities for the admin list ($p_i^{(a)}$) we use under-coverage estimates reported by the ONS (in Figure 7 of this source). These estimates are found by counting the number of Census usual residents that were linked to an admin source, but not included in the admin-based population estimates (ABPE) V3.0. This source provides counts for under-coverage for age-sex groups and only includes under-coverage for people aged 19-59 years old. These counts were converted to proportions and used as estimates of the admin list response probabilities. For ages outside that range, the response probabilities from the Census were used.

We include a hard-to-reach parameter in the simulation to represent a small proportion of individuals that have zero probability of being included on the Census or admin list. This proportion, represented by p_{htr}, factors in less common attributes that lead to under-coverage, such as homelessness and distrust in authorities. To simulate this, Np_{htr} individuals were randomly selected and their corresponding response probabilities $p_i^{(c)}$ and $p_i^{(a)}$ were set to zero. In the simulation we assume $p_{htr} < 0.05$.

With the response probabilities created for both sources, the list dependence can then be added. This will capture the conditional relationship between the two lists, i.e., if an individual is correctly captured on the admin list, how will that affect their likelihood of being captured correctly in the Census? There are many ways to simulate the list dependence, but for this simulation we use the odds ratio γ of whether an individual is on the Census and admin list. It is difficult to estimate a

realistic value of γ but it is assumed the association is positive, hence $\gamma > 1$. In the simulation we consider values between $1 < \gamma < 5$. Note that for individuals who were in non-responding households or were hard to reach, no dependence adjustment was made, so overall the observed odds ratio will be larger than the pre-specified value of γ. For details on how we integrated this dependence into the simulation see Appendix.

The two lists can now be selected using the derived values of p_h, $p_i^{(c)}$, $p_i^{(a)}$ and γ, which determine the under-coverage and list dependence in the simulation. This leaves the over-coverage that must still be integrated into the simulation.

2.3 Over-coverage

Under-coverage will occur whenever response probabilities are less than one, however over-coverage has yet to be integrated into the simulation. Over-coverage occurs in both the Census and the admin list but is expected to be a far greater problem in the latter. As such, we assume that over-coverage in the Census is non-existent, so over-coverage is integrated only into the admin list.

The over-coverage mechanism was introduced in the simulation by creating an entirely new and false synthetic population in the same way as the true synthetic population. Individuals in this false population will be selected based on an over-coverage rate, and they will then be randomly added to a true household on the admin list. These false additions represent the recently deceased or emigrated, as well as temporary residents registered elsewhere. The over-coverage rates are given within age-sex classes provided by the ONS (in Figures 4a and 4b). These over-coverage rates include all individuals in the ABPE V3.0 not linked to the Census minus all the Census non-response. This acts as a worst-case scenario of over-coverage with rates ranging between 0.04-0.14 across the different age-sex groups. It is reported that over-coverage will be improved by modifying the inclusion rules in the ABPE. Individuals in the false population are given over-coverage rates based on these figures and selected using a Bernoulli random variable. We include a simulation parameter $\delta \in (0,1)$ that reduces the over-coverage rate. If $\delta = 0$ then there is no over-coverage at all, and if $\delta = 1$ then the worst-case over-coverage rates will be used. For example, if $\delta = 0.5$, then this would imply that over-coverage has been improved by 50% of the currently believed worst-case scenario of over-coverage. Note that the over-coverage mechanism only includes individuals and so there is no over-coverage of households.

2.4 Linkage and estimation

With the two lists generated and over-coverage incorporated into the admin list, the two lists were then linked together to facilitate estimation. Linkage was done in two different ways depending on whether the estimator was using WCA or DSE.

For the WCA estimates, linkage is only at the household level. Linkage was done with the admin list acting as the frame, and the Census individuals were linked to this frame based on shared households. Hence, if individuals were missed in the Census, but not the admin list, then they would be added to the linked set if another individual were captured in the Census from their household. Once the Census and admin list were linked, the weights of the age-sex classes were calculated for class g like so:

$$w_g = \frac{Number\ of\ individuals\ in\ class\ g\ in\ admin\ list}{Number\ of\ individuals\ in\ class\ g\ in\ linked\ list}.$$

Using these weights, the WCA estimate of the population size of class g is:

$$\widehat{N}_g = w_g \hat{y}_g$$

where $\hat{y}_g = \sum_{i \in g} y_i$ and y_i is 1 or 0 depending on whether the i-th individual responded to the Census or not. From this the estimate of the total population size can then be found using $\widehat{N} = \sum_g \widehat{N}_g$.

For the DSE, the Census and admin list are linked at the individual level. The number of individuals in both lists are then counted, as well as the remainder who are solely in the Census or admin list respectively. These three counts are used to estimate the number of individuals missed by both lists, which can then be used to estimate the population size.

For both estimators in the simulation, we ignore data linkage errors that do occur in practice.

2.5 Parameter choice

The simulation study focuses on the effects of under- and over-coverage, and list dependence using parameters that determine the magnitude of these effects. For the purposes of this simulation study, we consider the parameters that determine under-coverage to be fixed. This includes the Census regression parameters and admin list under-coverage counts. Hence the parameters of greatest interest in the simulation study are the parameters that control the list dependence γ, over-coverage δ and hard-to-reach proportion p_{htr}. The three simulation parameters are displayed in Table 1. The default values represent the null value, which would be an ideal situation. So, if all three of these parameters are set to 0 then the under-coverage is the only factor that could lead to biased estimates. The simulation also incorporates three under-coverage "switches", one that ensures that the admin list has no under-coverage, a second that removes any Census non-response at the household level, and a third that removes any Census non-response at the individual level. Hence, if all three switches are on, there is no under-coverage in either list. Finally, the simulation allows for a simplified Census non-response with homogeneous response probabilities regardless of demographic. A fixed response probability can be set for all individuals, which removes any heterogeneity.

The true population size parameter N is also treated as fixed, which is set to the arbitrarily large size of 10,000. Varying N does not affect the relative bias or variance of the estimates.

Table 1 Simulation parameters

Parameter	Purpose	Default value	Considered range
γ	List dependence odds ratio between the Census and admin list.	1	1-5
δ	Over-coverage factor – representing the relative proportion of over-coverage individuals in admin list compared to worst-case scenario over-coverage, i.e., $\delta = 1$.	0	0-1
p_{htr}	Proportion of hard-to-reach individuals that will not be captured in either the Census or admin list.	0	0-0.05

Since it is difficult to know what the "true" parameters would be in reality, we vary them across a range of values that cover what is likely to be true. These ranges are specified in Table 1. One of the challenges of the simulation study is understanding how these three parameters affect the bias and variance of the weighting class estimates. To understand the causal effects, we consider scenarios where each parameter is varied one at a time, holding all else constant. We then introduce combinations of non-default values of parameters to get estimates of the bias and variance when

the parameters collectively act upon the simulation mechanisms. As a comparison, we also use DSE to get population estimates. Below is a list of all eight simulation scenarios (S):

S	Scenario description	Parameter values
1	Default values, representing a simple scenario where under-coverage exists, but list dependence and over-coverage are not present.	All default values: $(\gamma = 1, \delta = 0, p_{htr} = 0)$.
2	Increasing effect of list dependence.	$\gamma = 1, 2, 3, 4, 5$
3	Increasing effect of over-coverage on the admin list.	$\delta = 0, 0.2, \dots, 1.0$
4	Increasing effect of hard-to-reach proportion.	$p_{htr} = 0, 0.01, \dots, 0.05$
5	Same as Scenario 3 except no under-coverage for admin list.	$\delta = 0, 0.2, \dots, 1.0$
6	Default values except only household under-coverage in Census, hence no within-household under-coverage is present.	All default values
7	Default values except homogeneous within-household under-coverage set to 95%.	All default values
8	Default values except only within-household under-coverage in Census, hence no household under-coverage is present.	All default values

1000 simulations were generated for each scenario, with each iteration providing a different simulated population, census and administrative list. The 2.5th and 97.5th quantiles were used to represent a 95% interval. The width of the interval represents the variance of the estimates while the deviation of this interval away from N represents the bias. If the median estimate was approximately the same as N then the estimator is unbiased. Three estimators $\widehat{N}$ were assessed to estimate N:

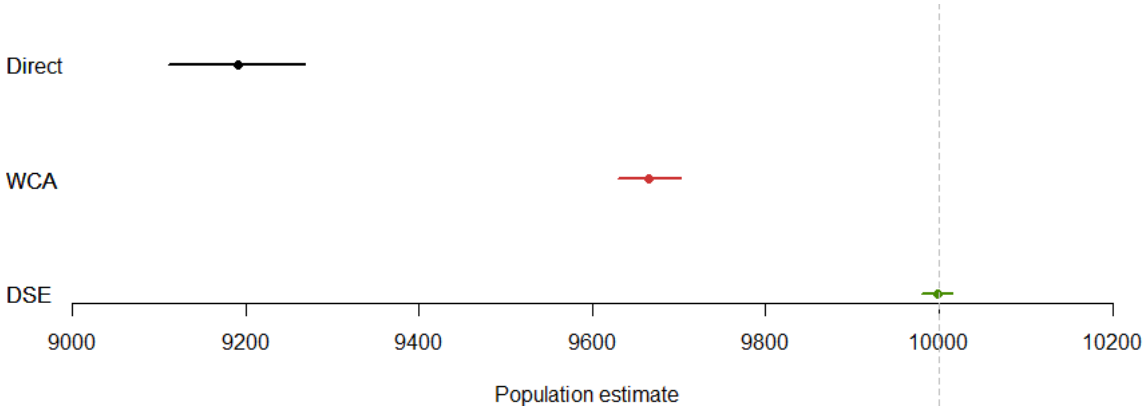
1. Dir – direct estimate based on the Census list ignoring the admin list. This is equivalent to the total number of Census responses.
2. WCA – this is the WCA estimate with linkage at the household level.
3. DSE – dual system estimator.

A further simulation scenario was then created, with varying rather than fixed parameters. This meant that each of the three parameters were assigned a distribution of feasible values. A Monte Carlo simulation was then conducted where each of the three parameters were randomly selected from the distribution and the simulations generated. This was repeated with 10,000 iterations and the distribution of estimates from each estimator recorded.

3 Results

3.1 Scenario 1 – default values

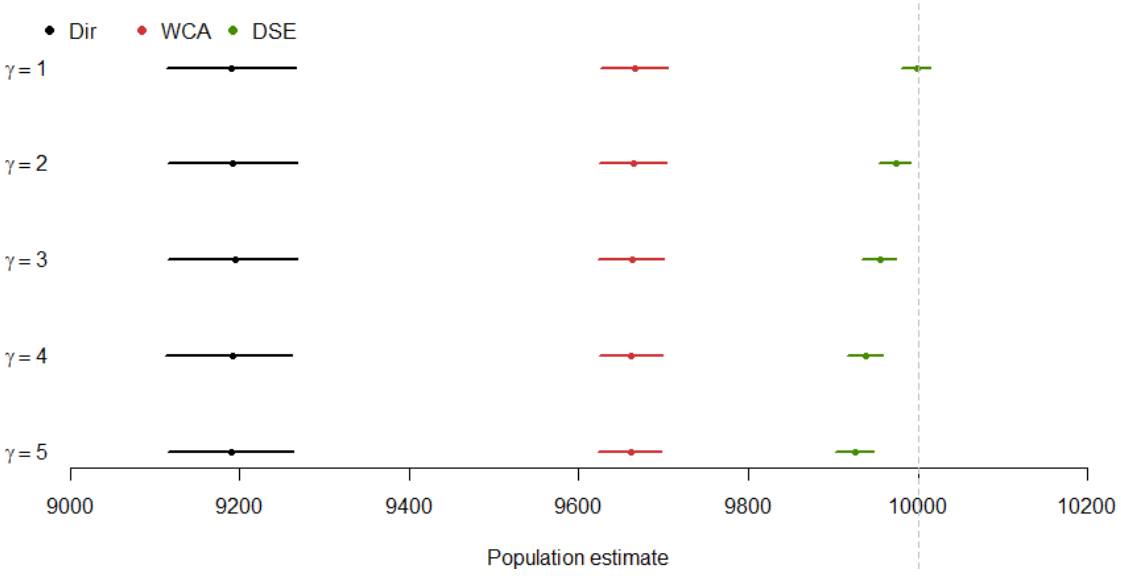
In this first scenario all default values are used meaning there is no added list dependence or over-coverage. The source of error comes from the household and individual under-coverage acting heterogeneously across age-sex classes.



The direct estimate shows the effect of the Census under-coverage derived from the regression models, which is approximately 0.92, slightly more under-coverage than the 2011 Census. The WCA estimates are negatively biased by approximately -3.4%, whereas DSE has a negligible negative bias. Furthermore, the variance of the estimates is greater for WCA than DSE.

3.2 Scenario 2 – list dependence

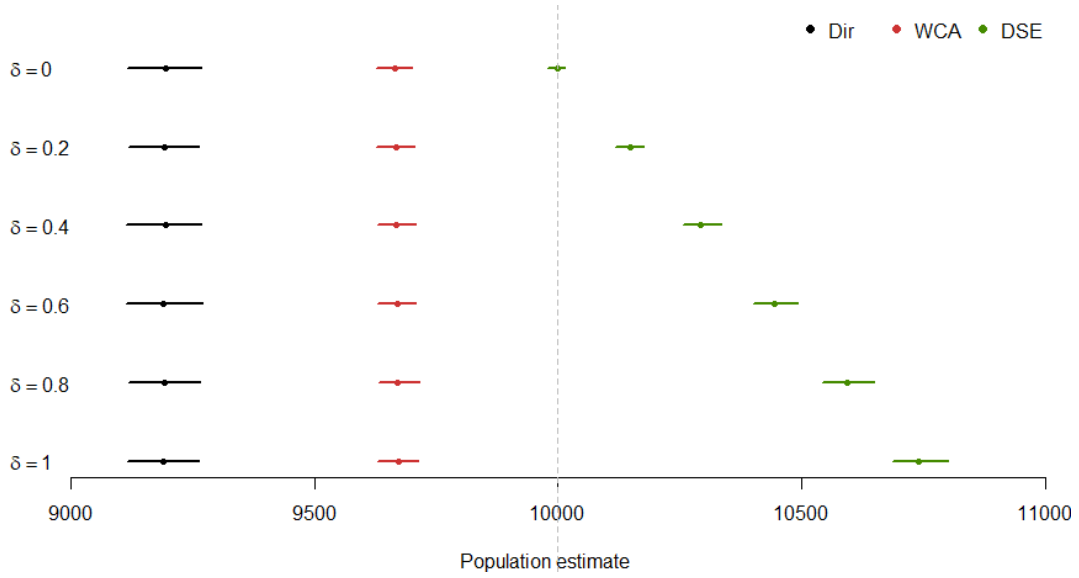
This scenario investigates the role of list dependence, with increasing odds ratios creating a stronger positive association between the response probabilities of the two lists. From the figure below it is clear that WCA estimates are more robust to list dependence than the DSE, showing no discernible difference across the values of γ. For DSE, the bias increases noticeably as γ increases, and the variance also. However, in reality the value of γ is unlikely to be very high, perhaps closer to 1 than 5. Regardless, even with a very high association, the bias and variance of the DSE are still considerably less than the WCA estimates.



3.3 Scenario 3 – over-coverage

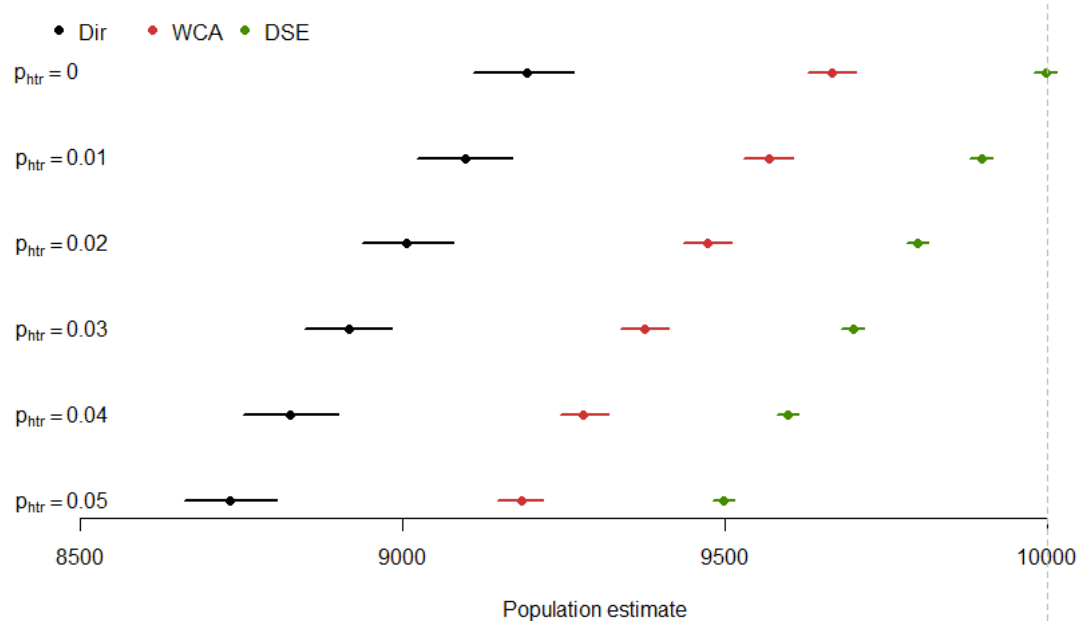
This scenario assesses the role of over-coverage on the estimates. The results in the figure show that WCA remains robust to the erroneous addition of individuals, unlike DSE. Under the worst-case scenario of over-coverage ($\delta = 1$) this leads to a bias of over 7%. WCA estimates are robust to over-coverage due to linking at the household level. This is because the over-coverage individuals are randomly assigned to the synthetic households, and these are not over-coverage households. This means that the over-coverage individuals will be counted in both the numerator and denominator (admin and linked list) of the weighting, meaning the estimates are mostly unaffected. For example,

suppose without over-coverage there were 100 individuals on the admin list in a given class and 90 found in the linked list, then the weight would be $w_g = \frac{100}{90} = 1.111$; now suppose there were 5% over-coverage, then the weight would change only slightly to $w_g = \frac{105}{95} = 1.105$. On close inspection of the WCA estimates, it appears there is a very minor increase in the estimates of N as over-coverage increases but for all intents and purposes this is negligible. This is in line with what would be expected.



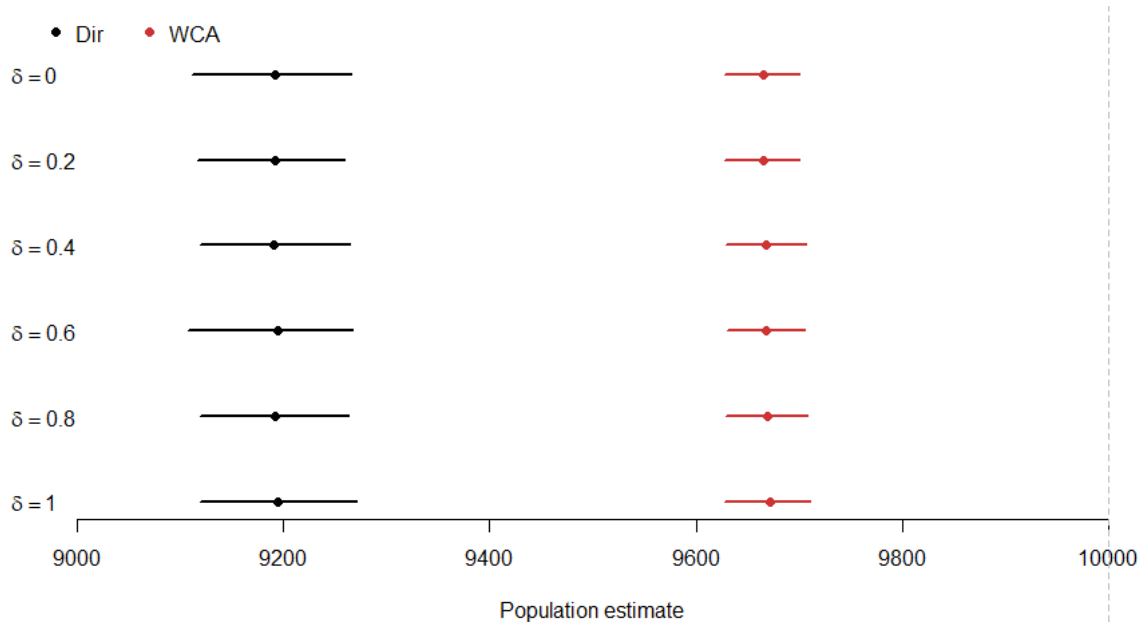
3.4 Scenario 4 – hard-to-reach individuals

This scenario assesses the effect of hard-to-reach individuals on the population size estimates. From the figure below it is clear that the hard-to-reach proportion negatively biases all the estimators. It does so at a uniform rate as the proportion increases, and also appears to affect all three estimates similarly. Therefore, hard-to-reach individuals are similarly problematic for the accuracy of both WCA and DSE.



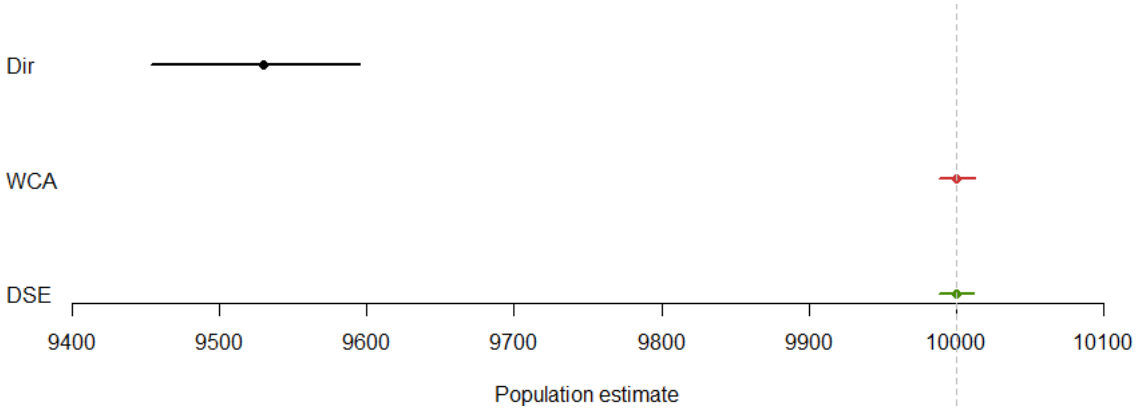
3.5 Scenario 5 – Complete admin list

In the previous four scenarios, the admin list had under-coverage. In this scenario, we assume the admin list has no under-coverage, hence a complete admin list is available. We vary the δ parameter in this scenario to observe the effects of added over-coverage. DSE could not be performed for this scenario because a complete list means that all individuals responding to the Census will also be found on the admin list. The results for this scenario are shown in the figure below. The primary finding is that the WCA estimates are not very different from scenario 3 where there was admin list under-coverage. This suggests that the large negative bias of the WCA estimates is not influenced by the quality of the admin list. Hence, neither under- nor over-coverage in the admin list are significant issues for WCA.



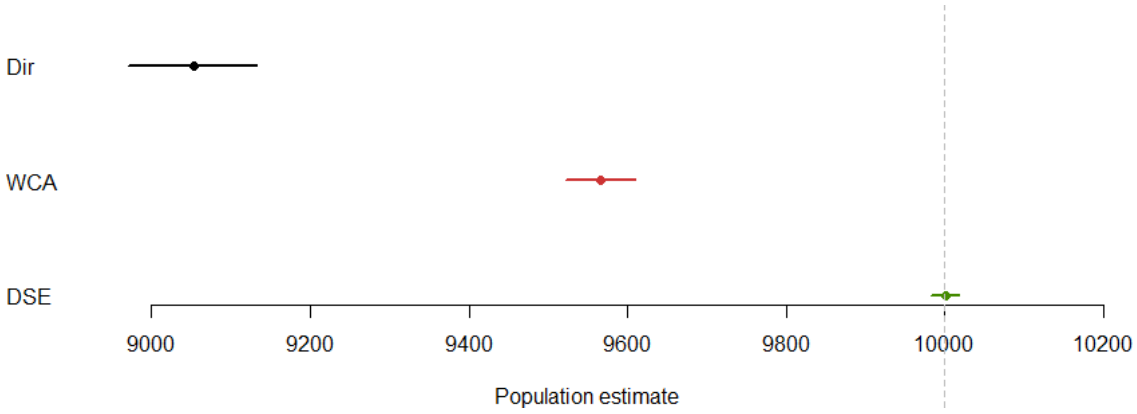
3.6 Scenario 6 – No Census within-household under-coverage (just household under-coverage)

In the previous scenario we found that the admin list is not a significant cause of the bias in the WCA estimates. This suggests that the problem must be occurring due to the under-coverage in the Census. The Census has two levels of under-coverage – at the individual and household level. In this scenario we assume under-coverage at the household level but not for individuals. Hence if one person responded to the Census, then all individuals in the same household would also have responded. Under-coverage in the admin list is also still present. The figure below shows the results of the simulation. The WCA estimates and DSE are both negligibly unbiased and have comparable variance. Although this means the DSE is similar to scenario 1, the WCA estimates are now completely without bias. This identifies the problem with WCA, which is the within-household under-coverage. It also shows that household under-coverage is not a significant problem for either estimator, under the conditions of the simulation.



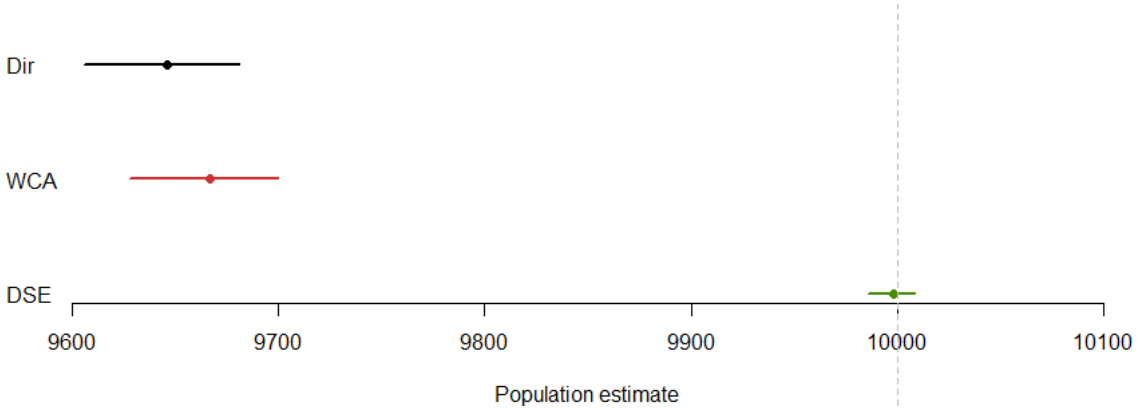
3.7 Scenario 7 – Homogeneous Census within-household under-coverage as well as household under-coverage

In the previous scenario we found when within-household under-coverage is removed, WCA is very accurate. Scenario 1 showed that WCA is not very accurate, with under-coverage similar to the 2011 Census. For this seventh scenario we introduce within-household under-coverage but ensure homogeneity in response rates by fixing all individuals from responding households to have a Census response rate of 0.95. This means the response rates do not vary by age-sex class. The results of the simulation for this scenario are shown in the figure below. Most notably, the WCA estimates are negatively biased by over 4% and DSE maintains its accuracy. This confirms that any within-household under-coverage adds bias to the WCA estimates, regardless of whether the response probability is homogeneous or not. The results appear to be very similar to the results in the first scenario, indicating that response heterogeneity is not a significant problem for either estimator.



3.8 Scenario 8 – No Census household under-coverage (just individual under-coverage)

Having now assessed the effects of within-household under-coverage we next assess the effect of household under-coverage. In this scenario we assume there is no household level non-response. Hence, under-coverage is only affected at the individual level. The results are shown in the figure below and show that without household-level under-coverage the WCA estimates are still negatively biased, and are comparable to the direct estimates. This is further evidence of the negative effects of within-household under-coverage on WCA. The variance of the WCA estimates is also large. The DSE performs well again, with a small but noticeable negative bias.

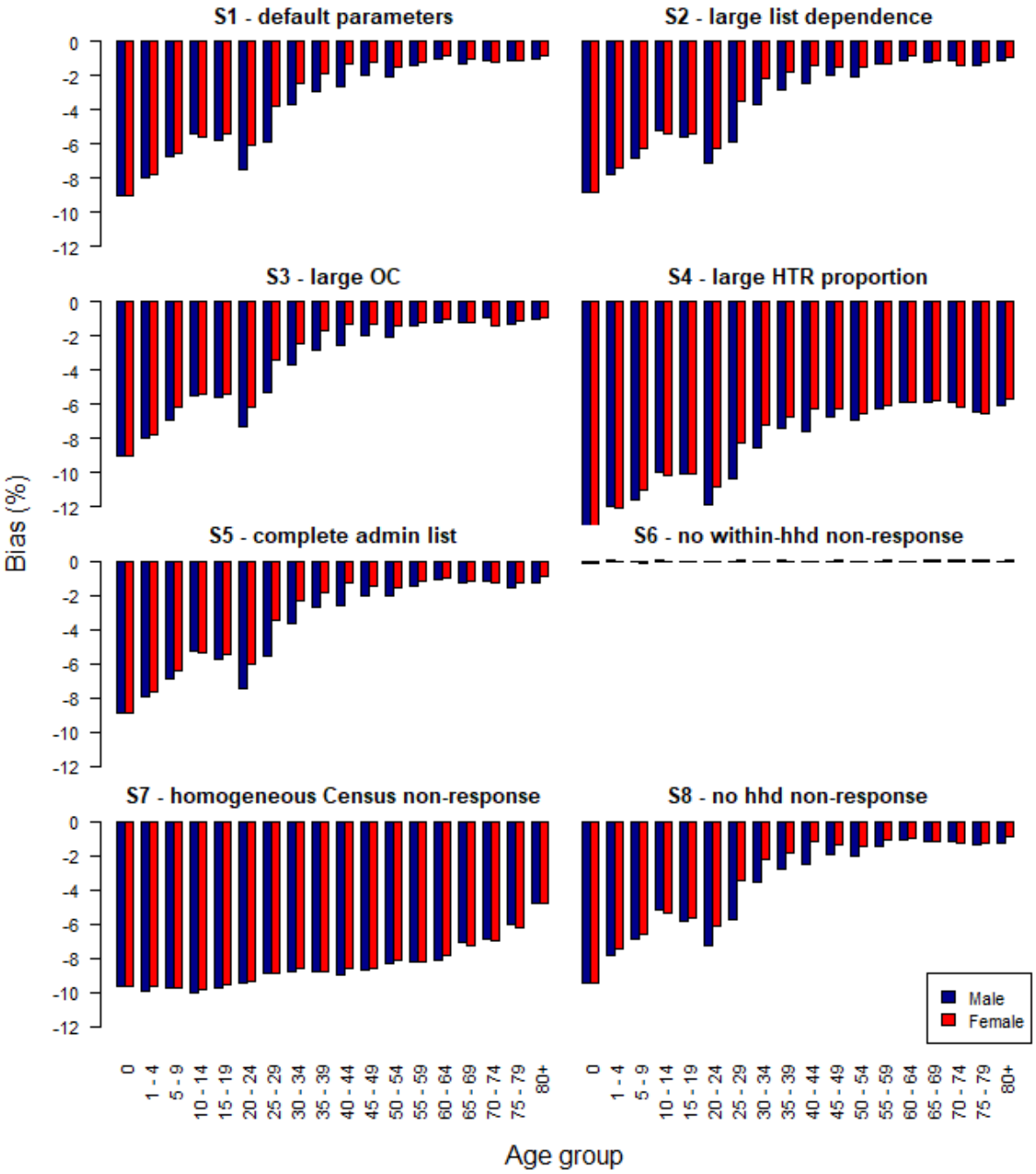


3.9 Bias within age-sex classes

To further investigate the bias of WCA estimates we examine the population size estimates within the age-sex classes. This is shown in the figure below, where the percentage bias, $100(\widehat{N}_g - N_g)/N_g$, for each class in each of the eight scenarios is reported. To remove some sampling error, we increased N to one million and did a single iteration since bias percentages did not vary much. For scenarios 2-5 that had multiple parameters above, we selected the most extreme value except for Scenario 5 which was the default value, i.e., for Scenario 2, $\gamma = 5$, for Scenario 3, $\delta = 1$, for Scenario 4, $p_{htr} = 0.05$ and for Scenario 5, $\delta = 0$.

The distribution of bias across the classes for Scenario 1 shows that generally younger age groups have larger bias. There is also a small spike in bias for the 20-24 age group, particularly for males. This bias distribution is very similar for Scenarios 2, 3, 5 and 8, highlighting how the effects of list dependence, over-coverage, having a complete admin list and having no household-level non-response all have minimal effect on the WCA estimates.

Scenario 4 differs to Scenario 1 as was expected. With the inclusion of a 5% proportion of individuals that are hard to reach occurring across all classes, the bias is uniformly affected by this. The bias percentages for Scenario 4 are approximately the same as Scenario 1, except with 5% more bias across all classes.



Scenario 7, with homogeneous within-household non-response shows bias that is more uniform across classes. This is to be expected, since heterogeneous under-coverage across classes has been shown to have an effect. What is interesting in this case, is that there is still a trend of reduced bias as age increases. This is due to the age-sex heterogeneity of single-person households. The older the age groups, the higher the proportions in single-person households. Single-person households are not affected by the over-representation of the age-sex counts in the linked list. This is because household-level linkage is the same as individual-level in these cases.

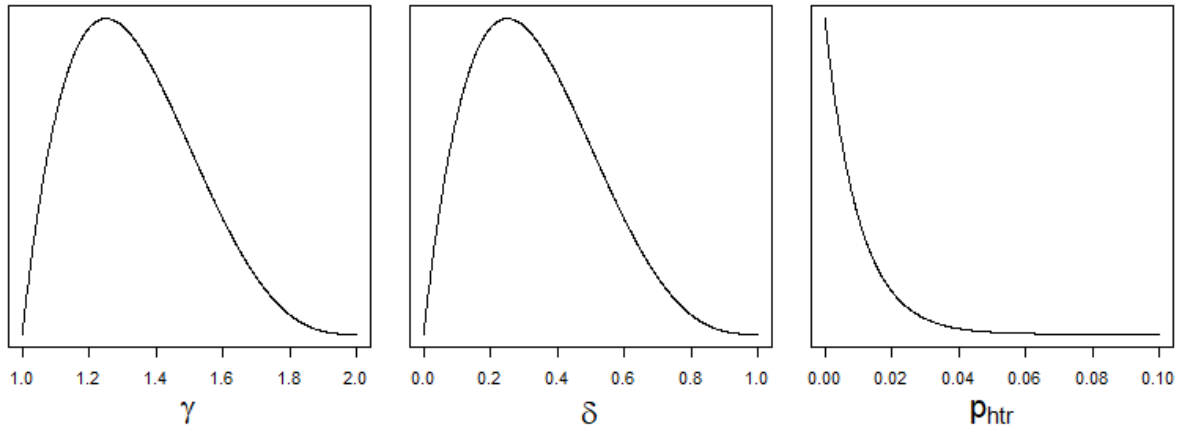
Finally, Scenario 6 shows an ideal situation where the bias for each class is approximately zero, which leads to the unbiased WCA estimates of N as shown in Section 3.6. It is useful to know that unbiased estimates of N were reached through unbiased estimates of N_g rather than biased estimates of N_g cancelling each other out.

3.10 Varying parameters

We have now assessed eight scenarios that have revealed some of the strengths and weaknesses of WCA estimation and DSE. In each of the scenarios one parameter is altered at a time to easily assess its effect. However, in reality, the factors that the parameters control will all collectively be present. So to evaluate the joint effect of all three parameters on the estimate of N we perform a Monte Carlo simulation where the unknown parameters are assigned distributions. For each iteration, a parameter value was randomly generated from a distribution. These distributions were subjectively selected such that they approximately represent the distribution of possible values. The distributions were:

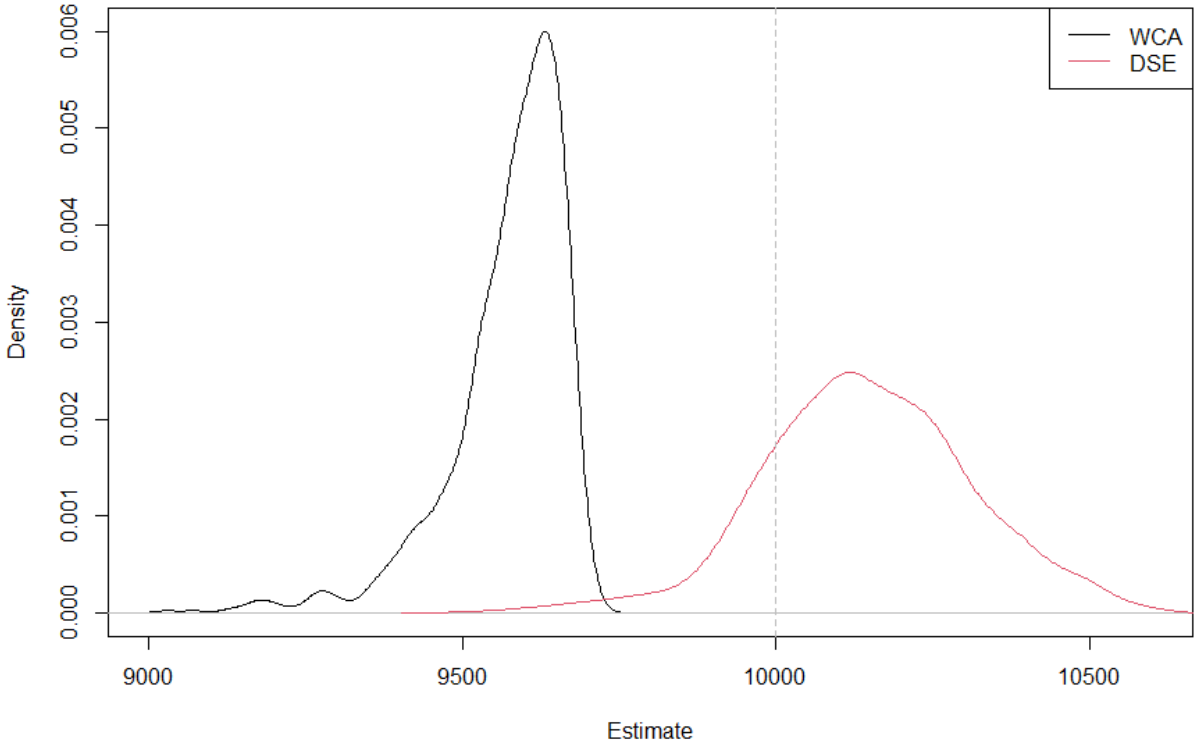
- $\gamma \sim 1 + Beta(2, 4)$
- $\delta \sim Beta(2, 4)$
- $p_{htr} \sim Beta(1, 100)$

and are shown in the figures below.



The resulting estimates from this Monte Carlo simulation are shown below based on 10,000 iterations. The WCA estimates are all significantly negatively biased, with a negative skewness caused by p_{htr}. The distribution for DSEs was more symmetric and positively biased on average. The variance of estimates is higher for DSE. This is because higher values of p_{htr} would cause an underestimate but higher values of δ lead to overestimates. The majority of the distribution of DSE being greater than N reflects the belief that over-coverage is likely to be a larger problem than hard-to-reach individuals.

Distributions of estimates for WCA and DSE using Monte Carlo simulation



4 Conclusions

The results of the simulation study have highlighted the strengths and weaknesses of WCA and DSE for population size estimates. The advantage of WCA is that it is almost completely robust to over-coverage and dependence between the Census and admin list. This is primarily due to the WCA estimator that links the two lists being based on household rather than individuals. Preliminary, but unreported findings showed that WCA based on individual-level linking leads to estimates very similar to DSE. Hence linking households has its benefits when over-coverage is known to be a significant problem. This comes at a cost though, primarily due to within-household under-coverage causing significant negative bias in the estimates.

To understand why within-household under-coverage leads to significant negative bias we consider how the weights are constructed. The weights are determined by the ratio of the counts in the admin list over the counts found in both lists, when linked by households. This weight will adjust for under-coverage in the Census if the admin list is close to a population frame. However, the consequence of linking at the household level is that the linked list includes all individuals on the admin list that are from a Census responding household. For example, a household of seven could be all included on the admin list, but on the Census list only one of the seven is present. Nevertheless, all seven individuals will be included on the list for both. As a consequence, the counts in the linked list will be higher than if the lists were linked on individuals. This leads to an over-representation in the counts on the linked list, leading to a deflated value of the weights. The reduced value of the weights does not compensate for the under-coverage leading to negative bias in the estimates. This was especially the case for age-sex classes that were more likely to be under-coverage in the Census, as well as those less likely to be in single-person households. Over-representation of linked counts cannot occur in single-person households, hence for these households linkage is equivalent to

individual level linkage. The reason that WCA works very well when there is under-coverage at the household level is because these over-represented individuals cannot exist, because there are no individuals that are present in the admin list but not the Census, yet share a household member in the Census. So because younger people are less likely to respond in the Census, and less likely to live on their own, they have the largest bias in population size estimates when using WCA. It is also worth noting that a WCA estimator was tested in the simulation study using age-sex by household size as the weighting classes, however there were only minor and non-significant improvements to the accuracy. Additional methods for defining more optimal classes could be explored in any future work.

DSE has opposing strengths and weaknesses compared to WCA. Under-coverage in both the Census and admin list does not lead to significantly biased population size estimates. This is even the case with under-coverage affecting each list differently, but also with some correlation in the response probabilities. However, over-coverage in the admin list directly leads to over-estimation. List dependence was a secondary factor in influencing estimates, but with considerably less impact. Even with very high dependence, unlikely to be seen in reality, the bias was less than 1%. With more realistic levels of dependence, i.e., $\gamma < 2$, this bias is estimated to be less than 0.3%. Although it is difficult to know what the dependence would be in reality, the simulation suggests it is unlikely to have a large effect on the estimates.

The simulation study highlighted how under- and over-coverage are the two biggest limiting factors in reaching accurate population size estimates. While WCA is robust to over-coverage in the admin list, DSE is robust to under-coverage in the Census. Hence if both methods were used the true population size would likely be found between the two estimates. For WCA to work well using age-sex classes, within-household under-coverage in the Census must be minimal. Although the simulated under-coverage is based on the 2011 Census it is actually slightly worse, and the 2021 Census may also have less under-coverage. Nevertheless, the simulation study suggests that within-household under-coverage would have to be negligible for WCA to be reasonably accurate.

5 Limitations

A simulation study is only useful when the assumptions made are in line with reality. Some of the assumptions in this simulation study were necessarily simplistic, which may affect the conclusions. The main assumptions of the simulation which may have limited the realities of the problem include:

- The synthetic population is representative of the 2021 population.
- The variation of characteristics within households is likely to be much smaller in reality compared to the synthetic population. In the synthetic population, households were formed randomly independent of age and sex, whereas in reality there are definitive patterns in age-sex characteristics within households. This may reduce the effect of household-level non-response.
- The 2021 Census will have similar household and within-household under-coverage to that of the 2011 Census, that vary solely on the variables used.
- There is no over-coverage in the Census.
- The under- and over-coverage in the admin list vary by age-sex groups and are comparable to the rates calculated based on ABPE V3.0.
- Over-coverage individuals in the admin list are found in existing households, and there are no households that are over-coverage.

If these assumptions are violated in reality, they could lead to different results.

Appendix - Generating list dependence using odds ratios

To create list dependence between the Census and admin list based on an odds ratio γ, $p_i^{(c)}$ and $p_i^{(a)}$ an adjustment must be made. Let C and A be the binary random variables indicating whether an individual is or is not (represented by a 1 or 0) on the Census (C) or admin (A) list. And let $p_i^{(ac)}$ be the joint probability associated with the i-th individual and C=c and A=A, e.g., $p_i^{(01)}$ indicates the probability of not being on the Census list but present on the admin list. The odds ratio γ can then be calculated as $\gamma = p_i^{(00)}p_i^{(11)}/p_i^{(01)}p_i^{(10)}$. If $p_i^{(c)}$ and $p_i^{(a)}$ are independent then there is no list dependence and $\gamma = 1$. If independent, then it is straightforward to calculate $p_i^{(ac)}$ using the marginal probabilities. To generate dependence, we add an adjustment factor ω to the joint probabilities like so:

$$p_i^{(00)} = 1 - p_i^{(c)} - p_i^{(a)} + \omega p_i^{(c)} p_i^{(a)}$$

$$p_i^{(01)} = p_i^{(a)} - \omega p_i^{(c)} p_i^{(a)}$$

$$p_i^{(10)} = p_i^{(c)} - \omega p_i^{(c)} p_i^{(a)}$$

$$p_i^{(11)} = \omega p_i^{(c)} p_i^{(a)} \ .$$

This ensures that the joint probabilities sum to one as required, and if $\omega = 1$ then independence is maintained. It can then be found that the relationship between γ and ω have a quadratic relationship hence a closed-formed solution exists, where ω equals:

$$\frac{1 - p_i^{(c)} - p_i^{(a)} + \gamma\left(p_i^{(c)} + p_i^{(a)}\right) - \sqrt{\left[p_i^{(c)} + p_i^{(a)} - 1 - \gamma\left(p_i^{(c)} + p_i^{(a)}\right)\right]^2 - 4p_i^{(c)}p_i^{(a)}\gamma(\gamma - 1)}}{2p_i^{(c)}p_i^{(a)}(\gamma - 1)}.$$

So, for each individual a value of ω can be found which can then be used to calculate the four joint probabilities. These probabilities can then be used to randomly generate the joint inclusion of an individual in each of the two lists. For example, for $p_i^{(11)} = 0.8$ the probability of (1,1) being selected is 0.8, leaving the probability of selecting (0,0), (0, 1) or (1,0) to be 0.2. The generation of the lists are therefore not done marginally, unless $\gamma = 1$.