The 2021 Census coverage survey: sample allocation strategy

Danielle Burke & Viktor Račinskij

Office for National Statistics

Titchfield, UK

February 20, 2020

**Note:** All results are subject to revision.

EXECUTIVE SUMMARY

*Motivation*

The way in which the census coverage survey sample is allocated is a crucial part of its overall design to enable coverage estimation. Changes to the 2021 Census mean that the sample allocation method needs reassessing. Data will be available quicker for the whole of England and Wales, which allows the coverage estimation strategy to use a generalized linear model approach. There will be borrowing of strength across areas, hence there may be less requirement for optimal allocation (the method in 2011). Furthermore, the 2021 Census will be online first. Therefore, it is more likely that the patterns of response may differ from those observed in the 2011 Census.

*Solution*

The idea for the sample allocation in 2021 is still to spread the sample across all local authorities reflecting anticipated variations by hard to count, but to guard against over-optimizing to the 2011 Census response patterns. In the context of the 2021 estimation strategy, an equivalent optimal allocation approach used in 2011 and a proportional allocation approach are considered, as well as a hybrid optimal-proportional allocation method. This hybrid method is a compromise strategy between the other two under consideration that ensures enough sample is allocated to all hard-to-count areas, but without allocating too closely to the patterns of response in 2011. Allocation methods are compared by observing variance of population total estimates in simulation studies. Additionally, the studies consider sampling fractions and maximum allocation constraints.

Overall, variances are small for all allocation methods with a generalized linear model approach. Generally, there are larger variances for proportional allocation. As desired, there is very little difference observed between the optimal and optimal-proportional approaches. Therefore, we recommend the use of the hybrid optimal-proportional allocation method for 2021.

We propose the use of a maximum constraint due to negligible difference with and without the constraint for the 2021 estimation approach. Constraints are particularly desirable to protect against operational difficulties of having extremely large sample in specific areas.

Lastly, the findings support the recommendation of earlier work to propose the smaller secondary sampling fraction of 0.25 of postcodes from each sampled output area (from 0.5 in 2011).

CONTENTS

[1. INTRODUCTION 2](#_Toc33175504)

[1.1. CCS 2021 design 2](#_Toc33175505)

[1.2. Sample allocation 3](#_Toc33175506)

[2. SAMPLE ALLOCATION METHOD IN CCS 2011 3](#_Toc33175507)

[3. SAMPLE ALLOCATION METHOD IN CCS 2021 4](#_Toc33175508)

[3.1. Simulation study: sampling design, sample allocation, and estimation 4](#_Toc33175509)

[3.2. Sample allocation methods under consideration 5](#_Toc33175510)

[3.3. Results from simulation study 6](#_Toc33175511)

[3.4. Discussion 30](#_Toc33175512)

# INTRODUCTION

The sample design of the Census Coverage Survey (CCS) is a critical part of the methodology for assessing coverage. The design sets out the sampling strategy, including the sample size, allocation, which units to sample, and stratification. This paper presents the findings so far on the allocation strategy for the 2021 CCS and corresponding sampling fractions.

## CCS 2021 design

The work to develop the CCS design has been through several iterations and reviews of both internal and external experts. The key principle behind the design is to use the information obtained in 2001 and 2011 to improve the sampling strategy, whilst at the same time guarding against over-optimising for the 2011 Census data. The work so far has culminated in the decisions outlined below (ONS, 2018).

* The CCS sample design is a two-stage approach in which output areas (OAs) are the primary sampling units (PSUs) and postcodes are the secondary (final) sampling units (SSUs).
* The stratification variables are the same as in 2011 with local authority (LA) as the primary geographical stratification, and the secondary stratification will be an updated version of the 5-level hard-to-count (HtC) index with a 40%, 40%, 10%, 8%, 2% distribution of the hard-to-count score. The updated version of the HtC index is the same index that is developed for the planning of the 2021 Census field follow-up.
* The overall sample size (i.e. the overall number of households) is the same as in 2011, which is approximately 335,000 households. However, for CCS fieldwork planning purposes the suggestion is to work with a range of (for example) ± 10,000 households, rather than a fixed number, because the sample size is fixed at the output area level, rather than a fixed number of households.

Outstanding decisions considered in this paper

* The sample allocation method has not yet been explored in conjunction with the proposed 2021 estimation strategy.
* Based on earlier work (Castaldo 2018), the current proposal for the 2021 CCS sampling fractions is to reduce the SSU sampling fraction from 0.5 (2011 level) to 0.25, which would be a less clustered design compared to 2011. This would, in turn, increase the number of PSUs in the sample. The earlier work involved a simulation study with four estimation areas and a ratio estimator. This paper will provide supporting evidence for the recommendation of the 2021 sampling fractions by evaluating the proposed strategy in conjunction with the proposed estimation and sample allocation methods.

## Sample allocation

The CCS sample allocation method is a crucial part of the design to enable coverage estimation (quality targets given in appendix A). Sample allocation in a stratified sample refers to how the sample is allocated to the different strata (i.e., sample size per stratum, *h* = 1, …, *H*) given a fixed overall size, *n*. For a stratified sample, two typical sample allocation methods are proportional allocation and optimal (also known as Neyman or disproportionate) allocation. Proportional allocation allocates the sample to strata, *h*, according to the relevant proportions in the population. Where the stratum sample proportion *fh* = *nh*/*n* is equal to the stratum population proportion, *Fh* = *Nh*/*N*, this implies a stratum *h* sample size is *nh* = *nFh*.

Optimal allocation allocates the sample to strata according to the stratum size and the variance within strata, *Sh2*, which captures the population behaviour with respect to whatever we are trying to measure (or with respect to a specific variable of interest); in the case of the CCS this is mainly census undercount (see Brown *et* *al*. 2011 for details). The sample size in stratum *h* is .

If the *Sh2* are more or less equal across all the strata, then this method of allocation is equivalent to proportional allocation. Since the variances are usually unknown, proportional allocation is probably the best allocation for increasing precision. In cases when the variances *Sh2* vary greatly, optimal allocation will give an estimator with smaller variance than proportional allocation (Lohr,1999). However, optimal allocation can introduce bias, and thus a trade-off between the reduction in variance and bias is often inevitable.

Section 2 outlines the CCS sample allocation method in 2011. Section 3 details the sample allocation strategies under consideration for the 2021 CCS sample with an outline of the simulation study, followed by results and discussion of findings.

# SAMPLE ALLOCATION METHOD IN CCS 2011

In 2011, for the estimation method of dual system, ratio, synthetic estimators, optimal allocation with the addition of some constraints was the selected allocation approach (Abbott, 2008b, Abbott and Brown, 2008, Brown, Abbott and Smith, 2011, ONS, 2010). A design variable based on the 2001 household imputation counts (which was the best measure available of census undercount) was used to determine the PSU (=output area) sample size within each stratum. The following constraints were also applied:

1. A minimum of one PSU per stratum (i.e. per local authority (LA) and hard-to-count (HtC) level)
2. A maximum of 60 PSUs in an LA (with a relaxed constraint in Leeds and Birmingham where the constraint was 1.5 and 2 times that of the other LAs, respectively).

The first constraint ensured that the sample was well spread across geography. Although this would not be enough for estimation purposes, it ensured that the sample of PSUs was ‘spread’ across the HtC index within each geographical stratum, providing some robustness. Additionally, the first constraint risked overrepresenting small strata in the final sample. Therefore, prior to allocation, to guard against over-allocating to small strata, the strata were collapsed across the HtC index if the number of PSUs within a stratum was less than 20. The second constraint prevented over-allocating the sample based on 2001 patterns. The design variable skewed the sample towards those strata where response was expected to be low or extremely variable, which tended to be urban and inner-city LAs where response was low in the 2001 Census.

The 2011 allocation method was different from that used in 2001 (predominantly proportional allocation). The 2001 allocation was relatively uniform across all areas and groups, whereas the 2011 allocation put more sample into areas where undercount was expected to be high based on 2001 Census data.

The 2011 approach to sample allocation worked well. For a fixed cost, the optimal allocation approach offered gains in the precision of the estimates at a national level, while smoothing the precision across local areas, when compared to both a proportional allocation and an allocation based on the 2001 methodology. The decision for the allocation method was supported by simulation studies conducted in preparation for the 2011 Census (Abbott, 2008b).

# SAMPLE ALLOCATION METHOD IN CCS 2021

Several changes between the 2011 and the 2021 Censuses, such as the data being available for the entire country quicker, permits the census coverage estimation strategy for 2021 to use a modelling approach using generalised linear models for a binary response. Therefore, the CCS sample allocation for 2021 approach requires reconsideration. The idea for the 2021 sample allocation method is still to spread the sample such that we can estimate coverage in all areas and for easy and hard to count areas. Another important consideration is that the chosen allocation method needs to guard against over-optimising to the 2011 Census response patterns because of changes in 2021; notably that it will be an online first census.

Using simulation studies (Račinskij, 2019), we present a comparison of the sample allocation methods under consideration alongside an evaluation of sampling fractions.

## Simulation study: sampling design, sample allocation, and estimation

An outline of the simulation study is briefly presented below. Full details of the data generation of the linked census and census coverage survey data are provided by Račinskij (2019) along with the estimation methods.

*Sampling design*

The CCS sample is a stratified (by hard-to-count index by local authority) two-stage cluster design with output areas as primary sampling units and postcodes as secondary sampling units.

Both the 2011 sampling fractions (~ 0.03 PSUs and 0.5 of SSUs from all sampled PSUs), and the proposed sampling fractions for 2021 (~ 0.05 of PSUs and 0.25 of SSUs from all sampled PSUs) are considered.

The sample size is the same as that in 2011 (335,000 households on average).

*Estimation*

Two methods are considered. The first is a logistic regression model approach (where a regression model approach is the proposed strategy for 2021) to estimate person totals with the following covariates: hard-to-count index, region, age-sex group, tenure, ethnicity, household relationship.

For comparison, the dual system, ratio, synthetic estimators (2011 approach) is considered where the ratio estimator is model-assisted.

Two CCS coverage rates can be considered: modelled CCS (based on the 2011 Census and CCS coverage rates) and perfect CCS (100% response rate). The work here assumed perfect CCS response.

## Sample allocation methods under consideration

Whilst we do not want undue reliance on the patterns of response in the 2011 Census data to allocate the sample, the choice of allocation method will involve some degree of trade-off between bias and variance in population totals. In proportional allocation, there is no design bias for the estimators under consideration, whereas this is not the case for disproportional allocation and logistic regression (both fixed and random effects models) estimators. The benefit of disproportional allocation is that it can increase precision compared to proportional allocation. However, if we are to estimate population totals in 2021 as expected using all the census data (rather than by estimation areas as we did in 2011), there will be borrowing of strength in estimation across areas, leading to the possibility that there is less gain in precision for optimal allocation compared to proportional when compared to the dual system, ratio, synthetic approach of 2011. This would allow the allocation strategy to be more conservative by spreading the sample more evenly across the areas. Therefore, both optimal and proportional allocation options are under consideration. Optimal allocation is derived by optimising for 2011 Census response patterns.

Due to the expected compromise between precision and bias, a third option of a hybrid optimal-proportional allocation is also considered. The idea is to allocate the sample optimally at hard-to-count level only (five strata), and then allocate the resulting five sample sizes proportionally to the local authorities within each hard-to-count index. This approach is potentially a middle ground between optimal and proportional allocation that ensures enough sample is allocated to each HtC stratum for estimation, while the proportional allocation within the HtC strata protects against allocating too closely to the 2011 patterns.

Beyond the statistical challenges, there are operational implications to consider. The optimal allocation approach could be most efficient in terms of reducing variance, however this choice could present operational challenges with very large sample in some hard to count areas. Proportional allocation would likely be the simplest approach for the fieldwork, with the hybrid approach potentially reducing some of the burden of the very large sample in localised areas.

Table 1: Sample design options for the 2021 CCS

|  |  |  |
| --- | --- | --- |
| **Option** | **Allocation method** | **PSU / SSU fraction** |
| 1 | Optimal allocation by LA x HtC index  Optimal allocation by LA x HtC index | 0.0241 / 0.5 |
| 2 | 0.0435 / 0.25 |
| 3 | Proportion allocation  Proportion allocation | 0.0241 / 0.5 |
| 4 | 0.0435 / 0.25 |
| 5 | Hybrid optimal-proportional allocation  Hybrid optimal-proportional allocation | 0.0241 / 0.5 |
| 6 | 0.0435 / 0.25 |

* + 1. **Constraints**

Furthermore, constraints on the allocation are considered. Like 2011, a minimum constraint of one PSU per stratum is applied to all options to make sure the sample is spread across all geography.

Optimal allocation options 1 and 2 are considered with both no upper bound constraint on the number of PSUs per stratum, and for equivalent 2011 maximum constraint. For the sampling fractions of PSU=0.0241, SSU=0.5, the maximum constraint is 60 PSUs in any LA (as in 2011). The maximum constraint is doubled to 120 PSUs for the SSU sampling fraction of 0.25.

To guard against over-allocating to small strata, the same constraint as 2011 was applied where, prior to allocation, the strata were collapsed across the HtC index if the number of PSUs within a stratum was less than 20.

Constraints are also beneficial for operational purposes to ensure that there are no exceptionally large samples allocated to local areas.

A comparison of the size of allocation to each stratum for optimal allocation compared to proportional and optimal-proportional is given in tables C1 and C2 in the appendix.

## Results from simulation study

Estimation is carried out for the entire England and Wales. Importantly, in this work we have not assessed bias in estimates because bias is introduced in optimal allocation for logistic regression estimation. Therefore, the quality of estimates is assessed by considering only the relative standard errors (RSE) for the local authority by hard-to-count person totals, and local authority person totals. At this stage the results are based on 100 simulation iterations.

* + 1. **Relative standard errors for LA / HtC totals (total of 848 points)**

*Comparison of allocation methods*

Figure 1 shows the log RSEs for all LA x HtC strata estimates for the three allocation methods with sampling fractions of PSU=0.0435 (approximately 8327 output areas) and SSU=0.25 (options 2, 4 and 6 in Table 1) and estimation by logistic regression model. Figure 2 presents the same results by HtC index (equivalent box plots shown in Figure 3).

One can expect that the RSE will be affected by the size of the estimation domain. Overall, the RSEs are small and for the smallest RSEs, the three methods are very similar. For some of the larger RSEs there are some differences with those for proportional allocation being slightly larger. The RSEs for the optimal and hybrid optimal-proportional allocation methods are very similar.

The size of RSE and difference across allocation methods vary by hard-to-count index. For LA x HtC 1 and 2 totals, the RSE are the smallest and similar across all allocation options. There are differences across allocation method for LA x HtC 3 totals, but overall the range of values are similar across all three options. It is for LA x HtC 4 and 5 totals that there are more obvious differences in RSE across allocation methods, with larger RSE for proportional allocation. Those for optimal and optimal-proportional allocation methods differ, however overall there is little difference in the size of RSE.

1. b)

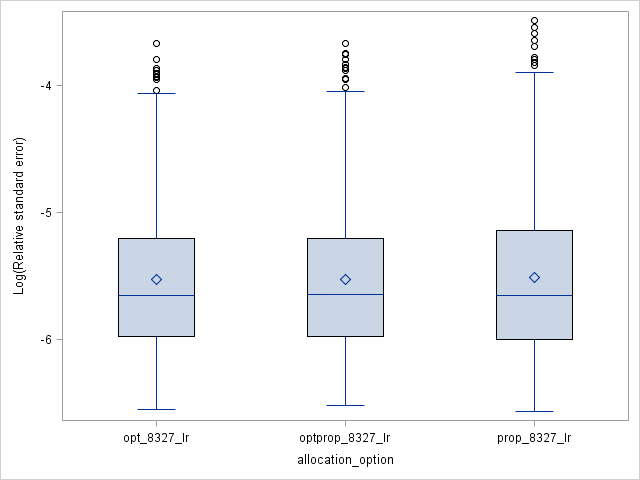
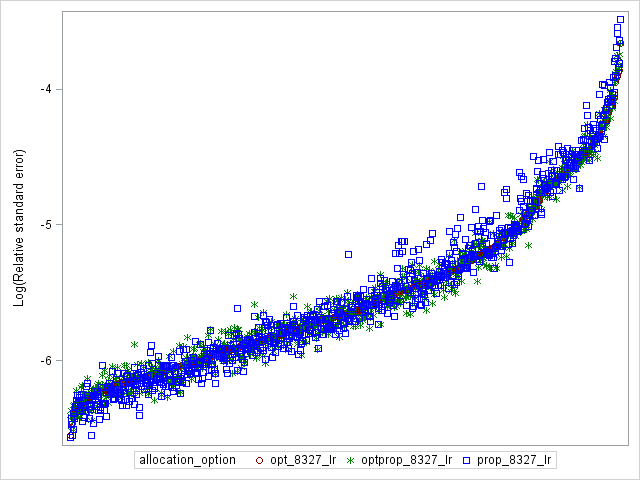


Figure 1: : log(RSE) for all LA x HtC totals for optimal (opt\_8327\_lr), hybrid optimal/proportional allocation (optprop\_8327\_lr), and proportional (prop\_8327\_lr), and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size in optimal allocation with PSU sampling fraction 0.0435, SSU sampling fraction 0.25; (b) corresponding box plot.

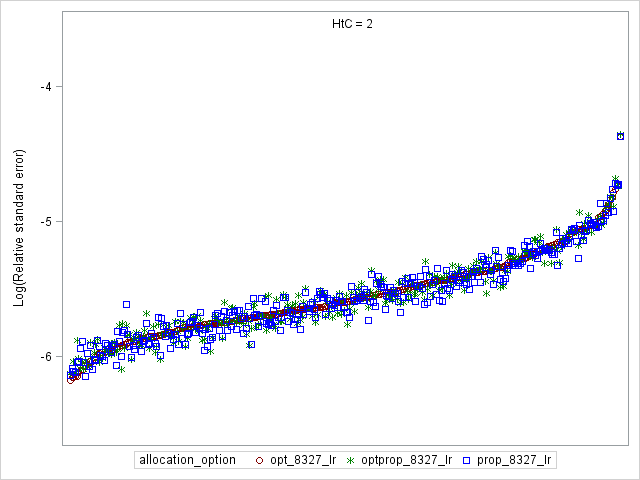
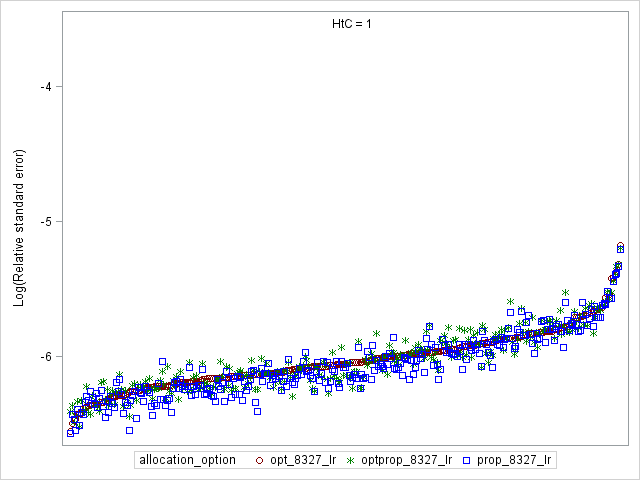
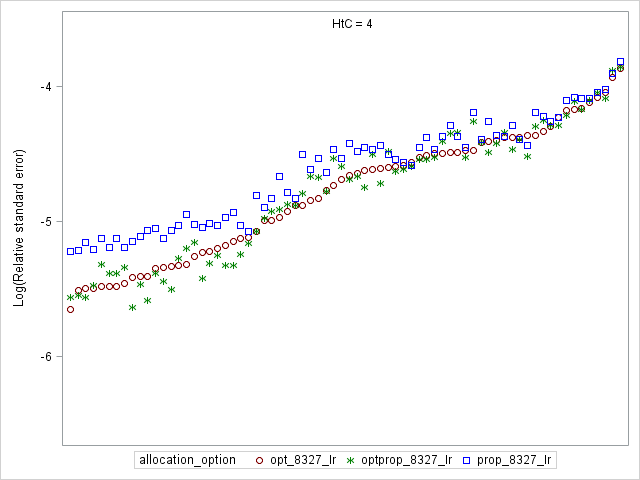
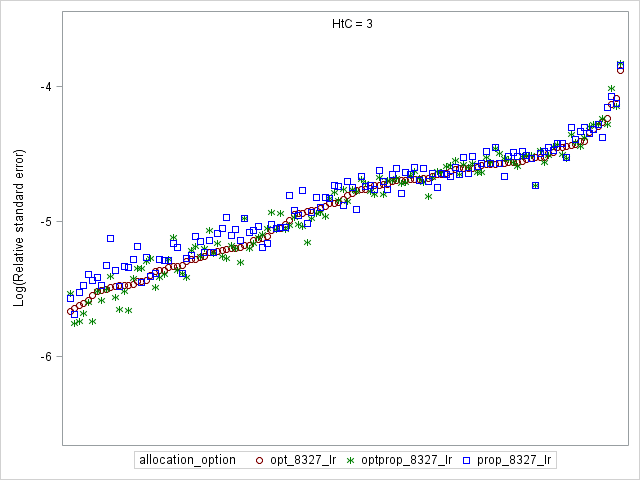
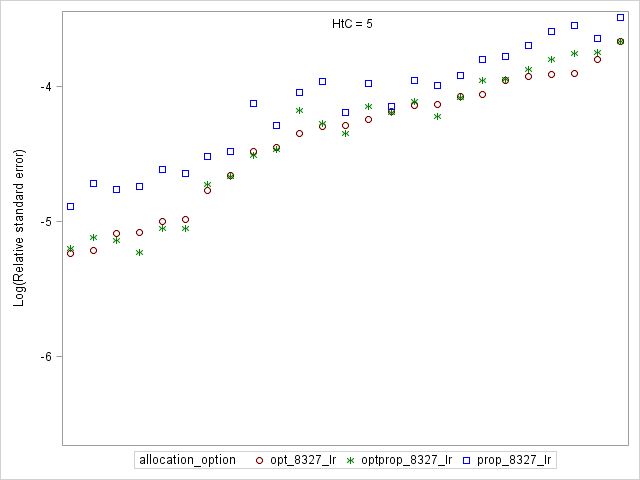
  

Figure 2: Scatter plots of log(RSE) for LA x HtC totals for optimal (opt\_8327\_lr), hybrid optimal/proportional allocation (optprop\_8327\_lr), and proportional (prop\_8327\_lr), ordered by log(RSE) size (optimal allocation) and displayed by HtC index category. Logistic regression model estimation.

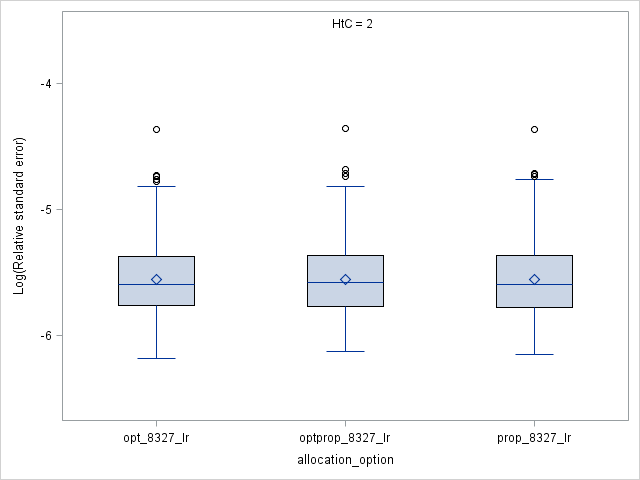
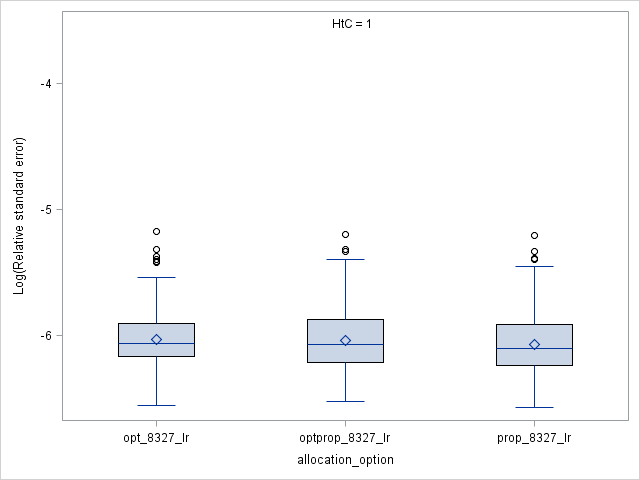
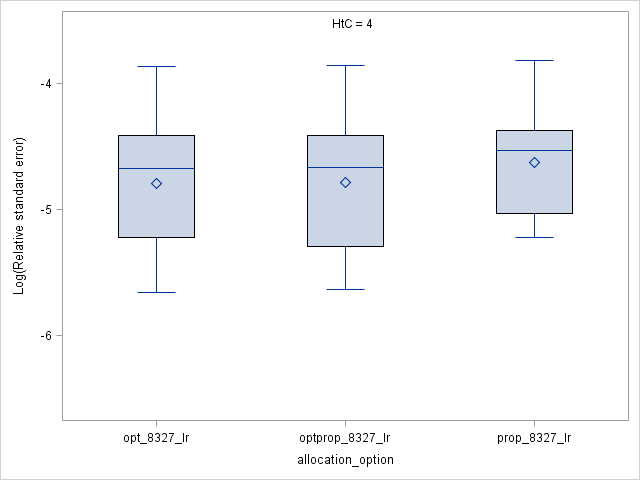
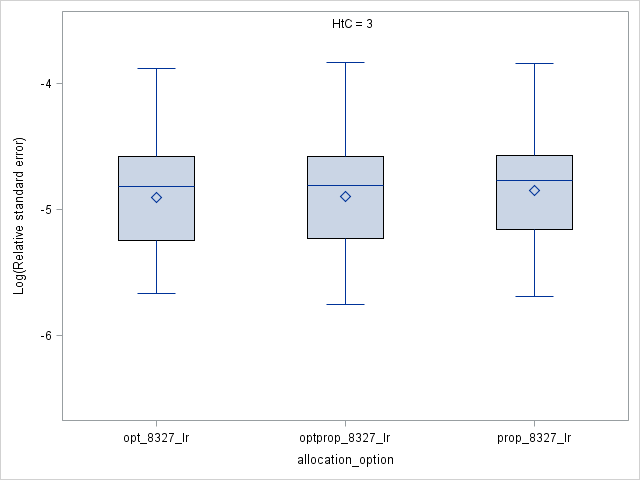
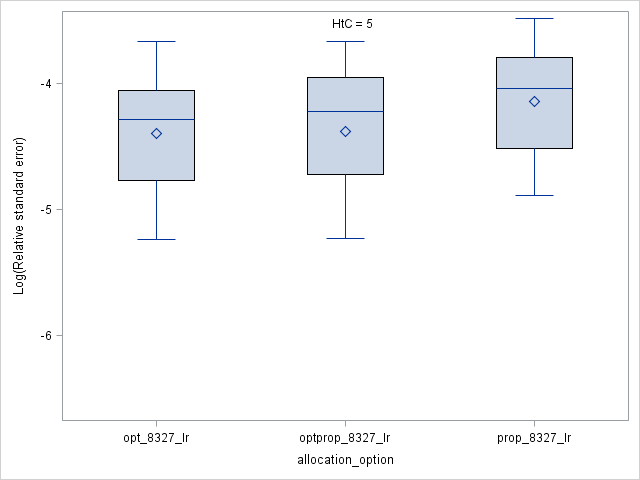
  

Figure 3: Box plots of log(RSE) for LA x HtC totals for optimal (opt\_8327\_lr), proportional (prop\_8327\_lr), and hybrid optimal/proportional allocation (optprop\_8327\_lr) and displayed by HtC index category. Logistic regression model estimation.

*Comparison of sampling fractions*

This section examines differences in precision between sampling fraction strategies by considering each sample allocation option independently. First, we consider optimal allocation.

Optimal allocation

Figure 4 shows the RSE for all LA x HtC strata estimates with logistic regression estimation to show a comparison of the 2011 and 2021 sampling fractions. Figure 5 presents the same results by HtC index (equivalent box plots shown in Figure 6).

1. b)

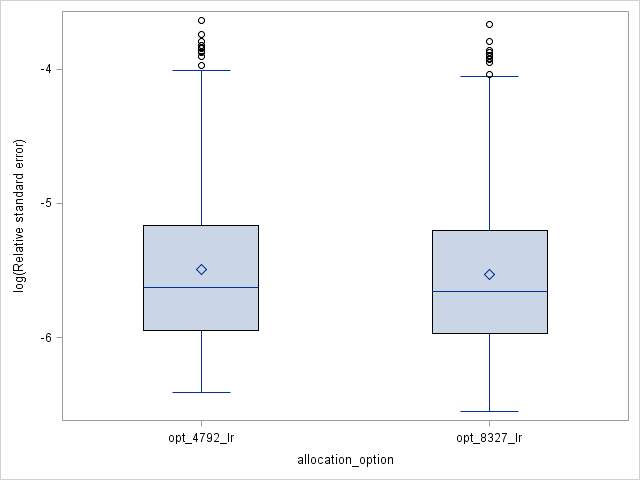
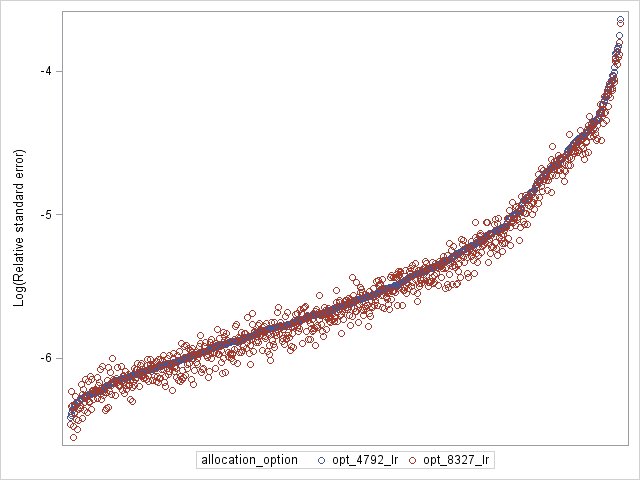


Figure 4: Log(RSE) for all LA x HtC totals for optimal allocation options 1 and 2, and logistic regression model estimation. Scatter plot (a) ordered by RSE size with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (opt\_4792\_lr). PSU sampling fraction 0.0435, SSU sampling fraction 0.25 is opt\_8327\_lr; (b) corresponding box plot.

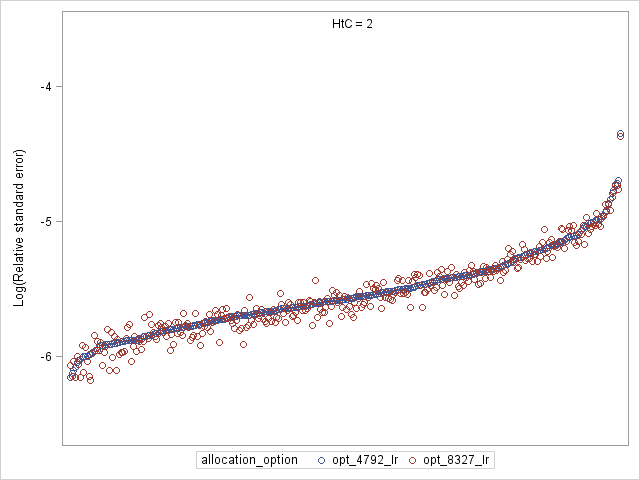
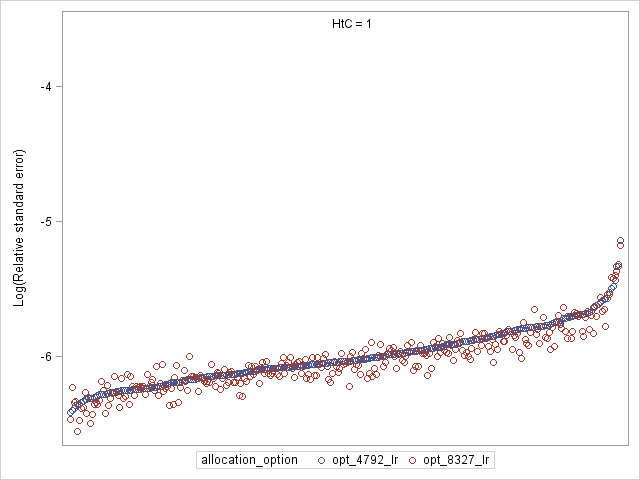
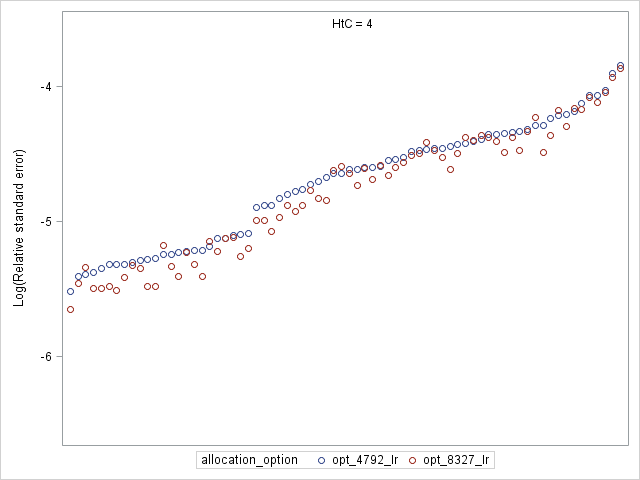
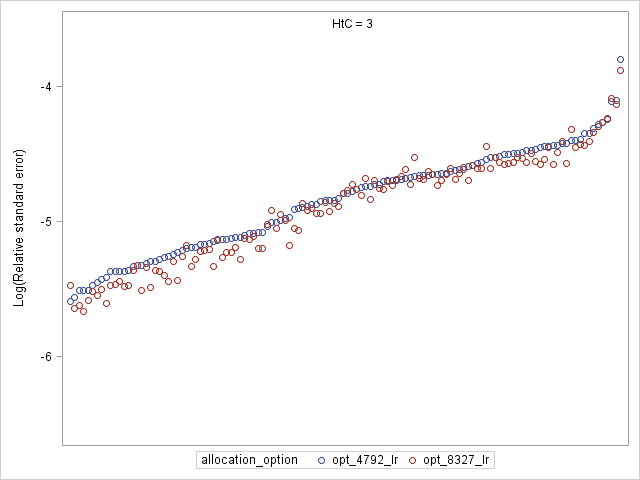
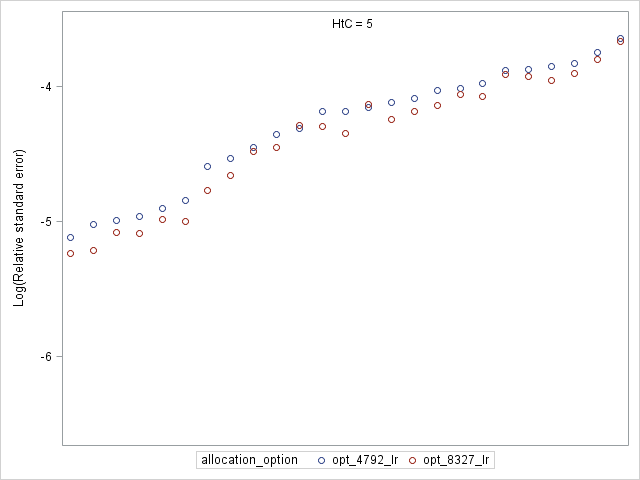
  

Figure 5: Scatter plots of log(RSE) for LA x HtC totals for optimal allocation options 1 and 2, and logistic regression model estimation. Scatter plots with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (opt\_4792\_lr) and PSU sampling fraction 0.0435, SSU sampling fraction 0.25 (opt\_8327\_lr). Ordered by log(RSE) size for opt\_4792\_lr and displayed by HtC index category.

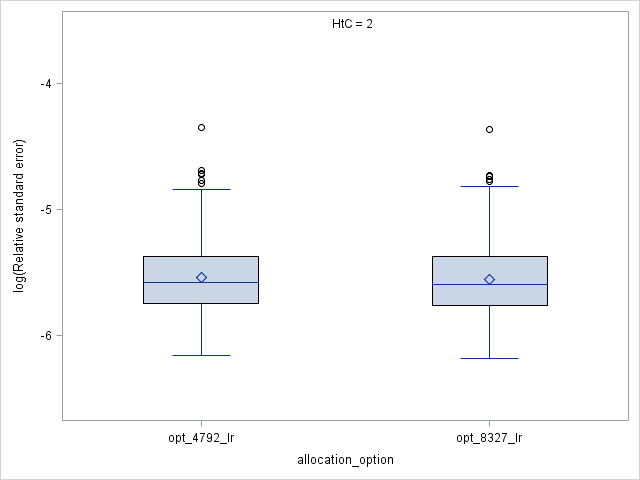
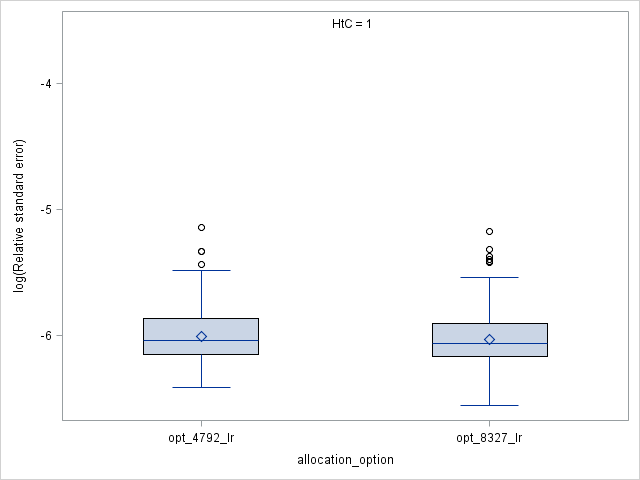
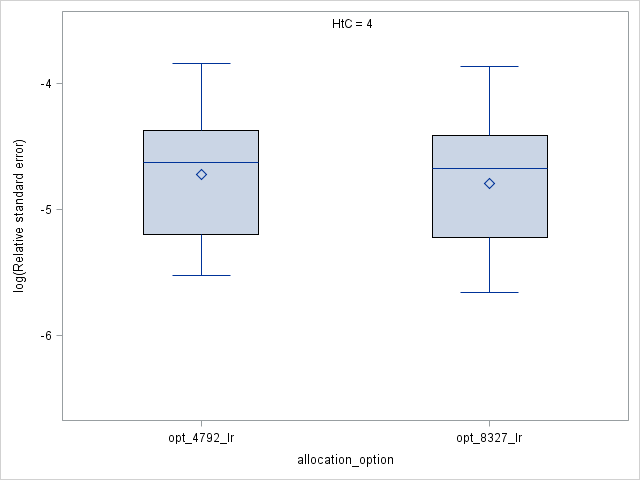
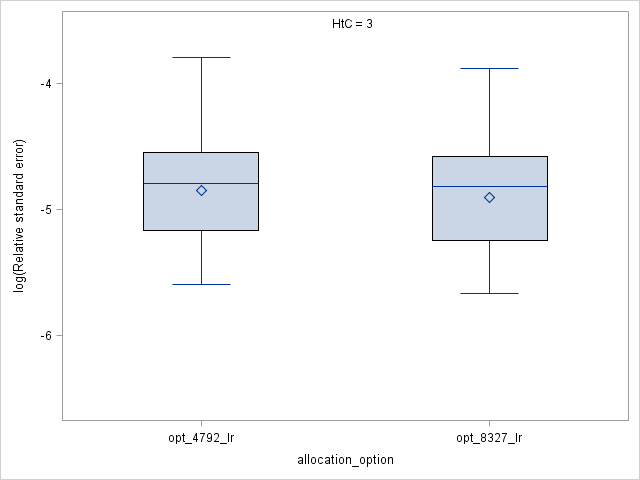
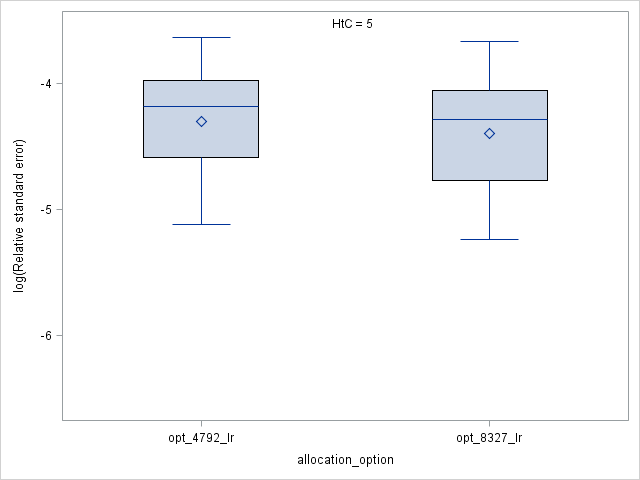
  

Figure 6: Box plots of log(RSE) for LA x HtC totals for optimal allocation options 1 and 2. PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (opt\_4792\_lr) and PSU sampling fraction 0.0435, SSU sampling fraction 0.25 (opt\_8327\_lr). Displayed by HtC index category. Logistic regression model estimation.

Overall, there are slightly lower RSE for the less clustered design (option 2 with sampling fractions PSU=0.044 and SSU=0.25). As with the comparison of allocation strategies, there appears to be negligible difference in RSE for the LA x HtC 1 and 2 strata. Differences between sampling fractions become more pronounced as the HtC index increases with most RSE being lower for the less clustered design. The differences in RSE are largest for LA / HtC 5 strata estimates.

Proportional allocation

Figure 7, Figure 8, and Figure 9 display the equivalent figures as those in Figure 4, Figure 5, and Figure 6 for proportional allocation.

1. b)

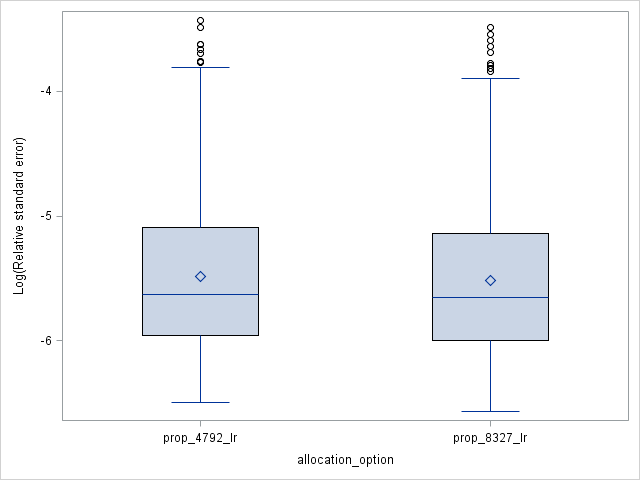
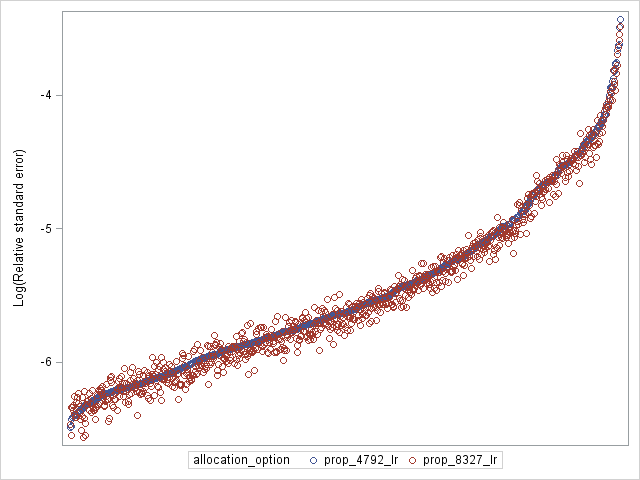


Figure 7: Log(RSE) for all LA x HtC totals for proportional allocation options 3 and 4, and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (prop\_4792\_lr). PSU sampling fraction 0.0435, SSU sampling fraction 0.25 is prop\_8327\_lr; (b) corresponding box plot.

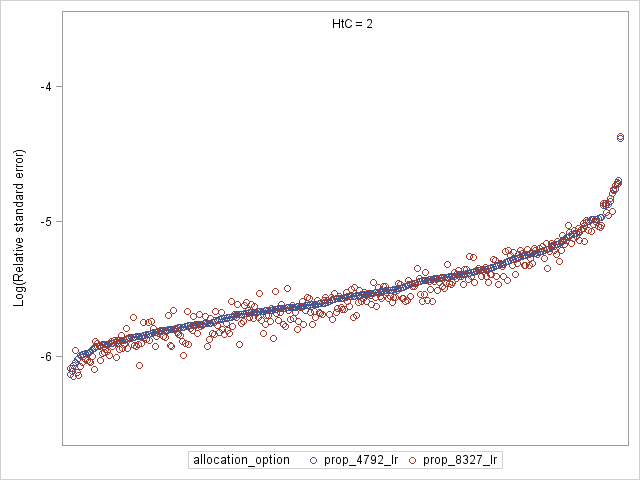
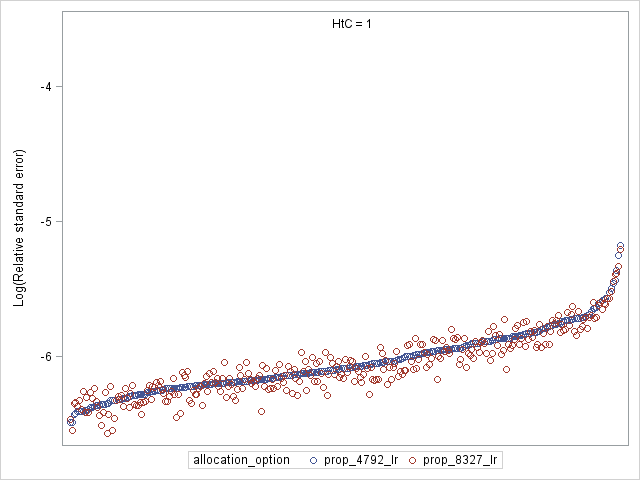
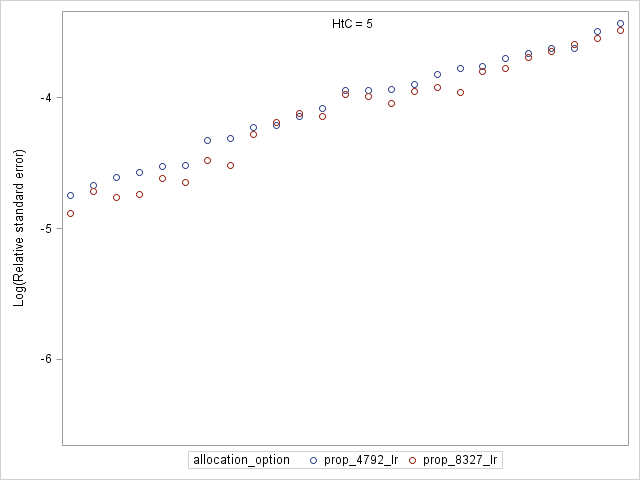
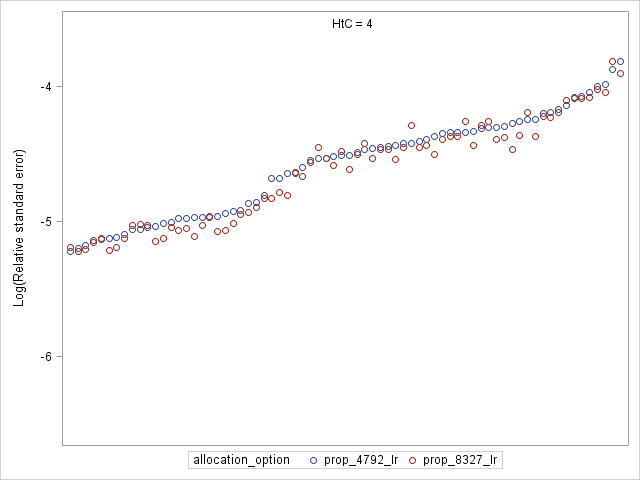
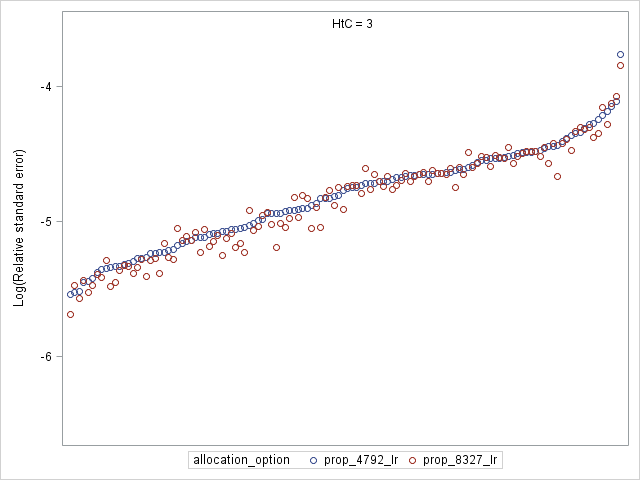
 

Figure 8: Scatter plots of RSE for LA x HtC totals for proportional allocation options 3 and 4, and logistic regression model estimation. Scatter plots with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (prop\_4792\_lr) and PSU sampling fraction 0.0435, SSU sampling fraction 0.25 (prop\_8327\_lr). Ordered by log(RSE) size for prop\_4792\_lr and displayed by HtC index category.

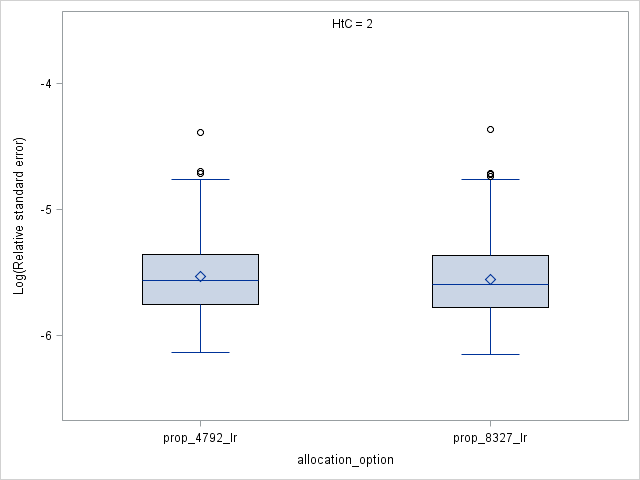
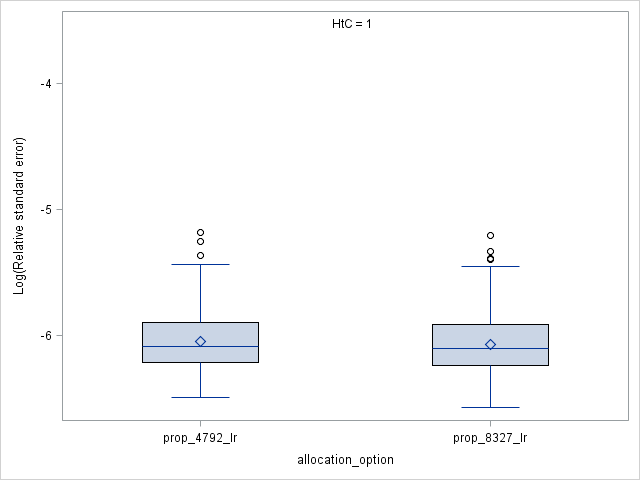
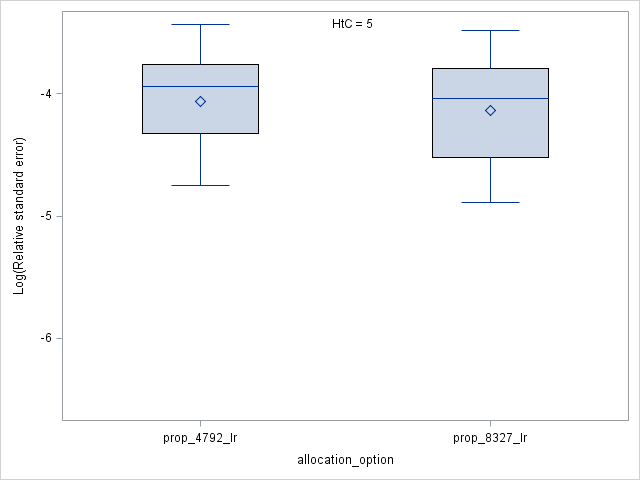
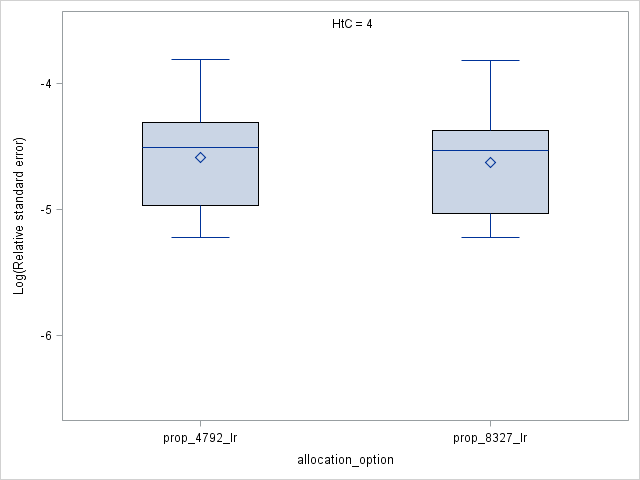
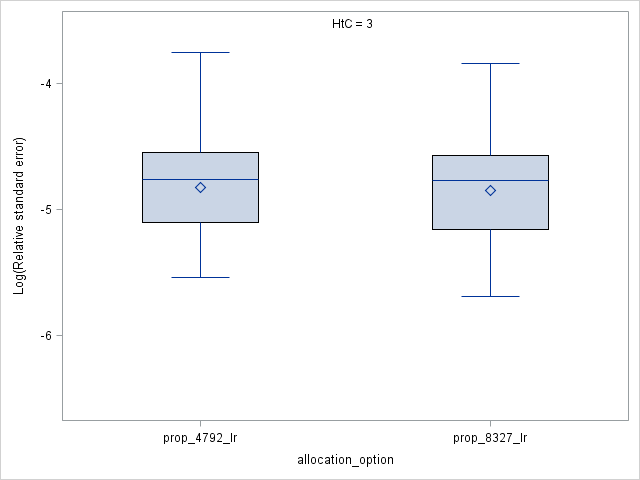
 

Figure 9: Box plots of log(RSE) for LA x HtC totals for proportional allocation options 3 and 4. PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (prop\_4792\_lr) and PSU sampling fraction 0.0435, SSU sampling fraction 0.25 (prop\_8327\_lr). Displayed by HtC index category. Logistic regression model estimation.

As with optimal allocation, overall there are slightly lower RSEs for the less clustered design (PSU=0.0435 and SSU=0.25) however the differences are smaller compared to those with optimal allocation. For LA x HtC 5 strata, the RSEs are consistently lower.

Hybrid optimal-proportional allocation

Figure 10, Figure 11 and Figure 12 display the equivalent figures for the hybrid optimal-proportional allocation as those shown for optimal and proportional allocation above.

1. b)

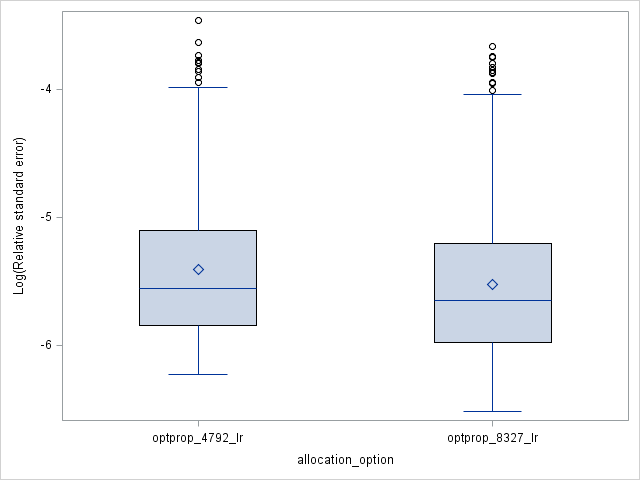
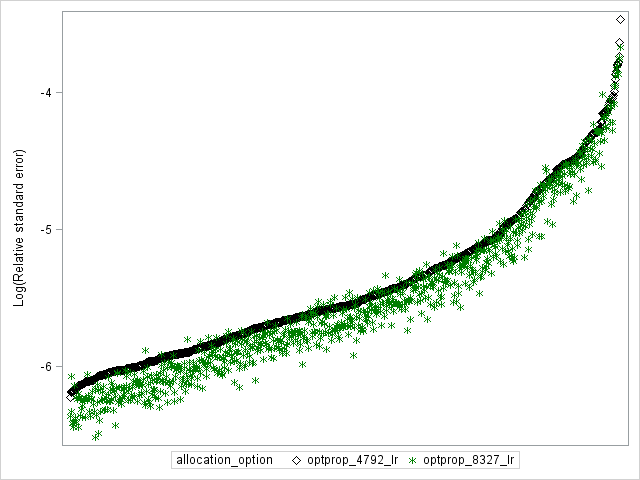


Figure 10: log(RSE) for all LA x HtC totals for optimal-proportional allocation options 5 and 6, and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (optprop\_4792\_lr). PSU sampling fraction 0.0435, SSU sampling fraction 0.25 is optprop\_8327\_lr; (b) corresponding box plot.

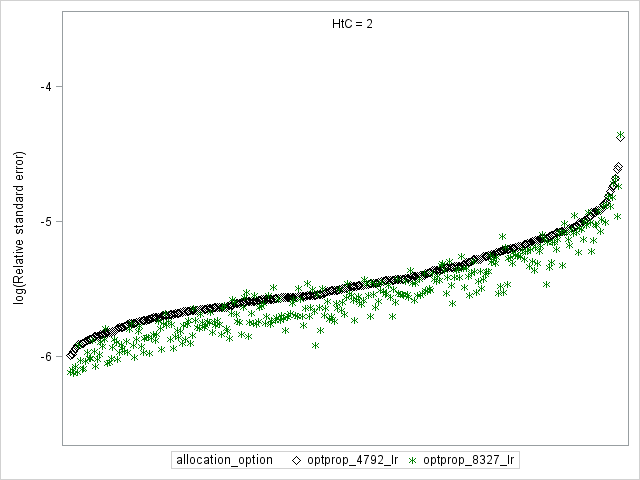
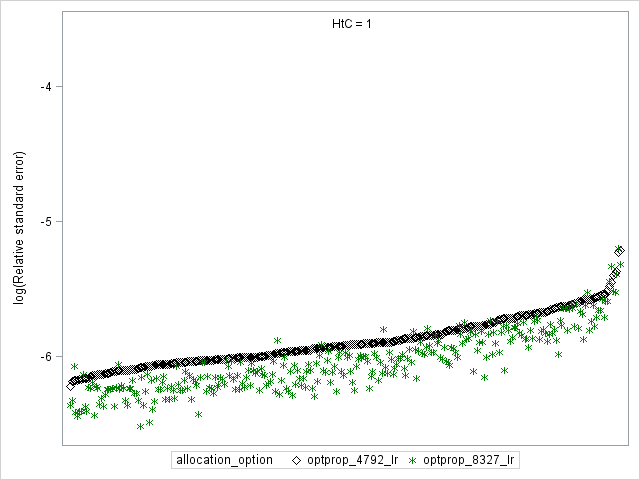
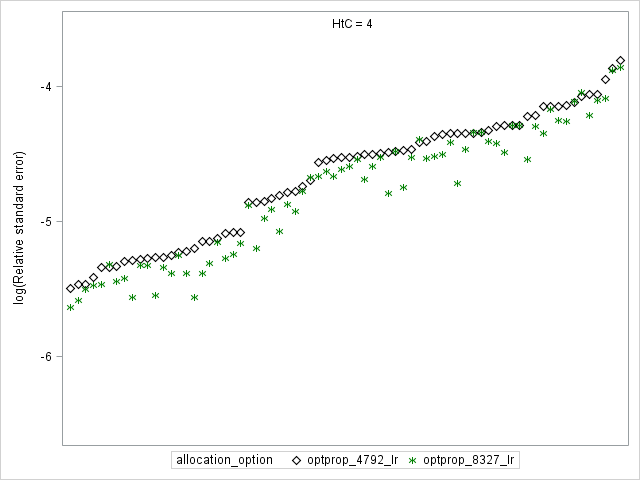
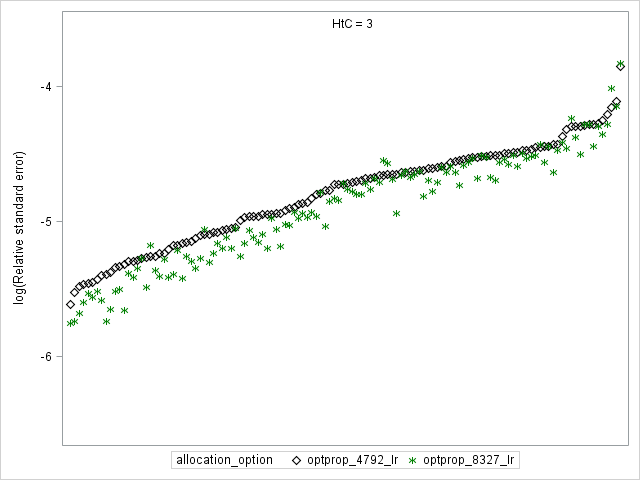
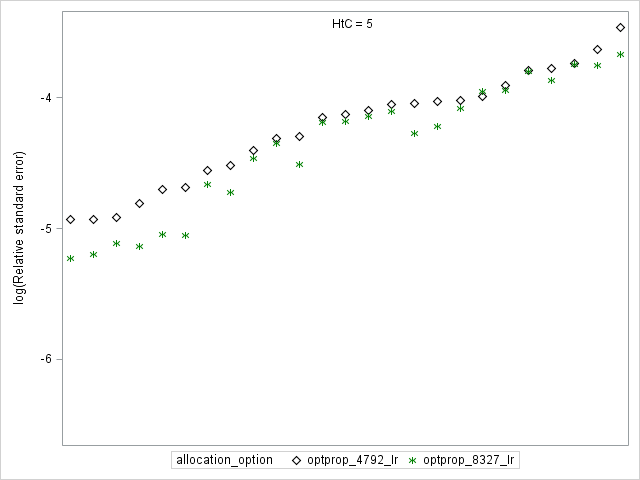
  

Figure 11: Scatter plots of log(RSE) for LA x HtC totals for optimal-proportional allocation options 5 and 6, and logistic regression model estimation. Scatter plots with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (optprop\_4792\_lr) and PSU sampling fraction 0.0435, SSU sampling fraction 0.25 (optprop\_8327\_lr). Ordered by log(RSE) size for optprop\_4792\_lr and displayed by HtC index category.

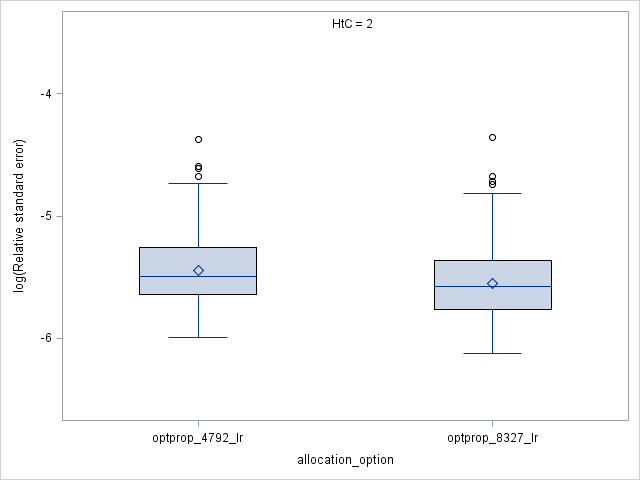
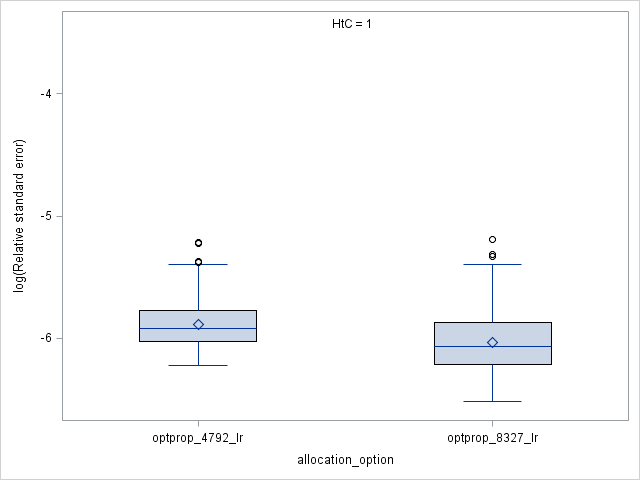
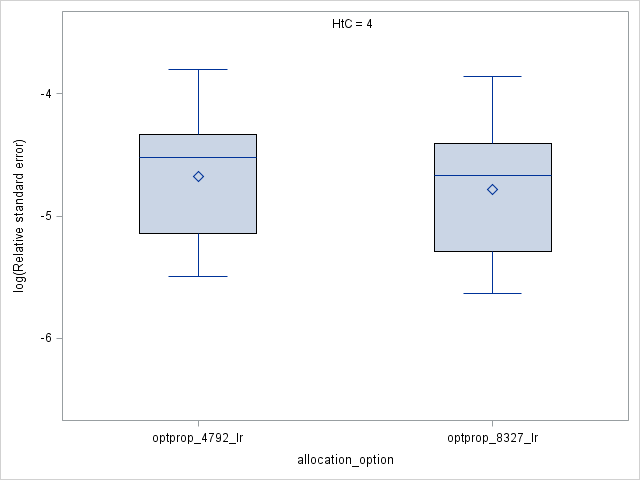
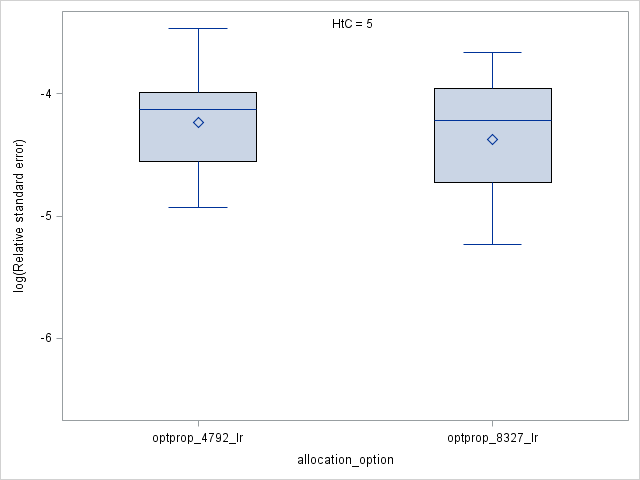
  

Figure 12: Box plots of log(RSE) for LA x HtC totals for optimal-proportional allocation options 5 and 6. PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (optprop\_4792\_lr) and PSU sampling fraction 0.0435, SSU sampling fraction 0.25 (optprop\_8327\_lr). Displayed by HtC index category. Logistic regression model estimation.

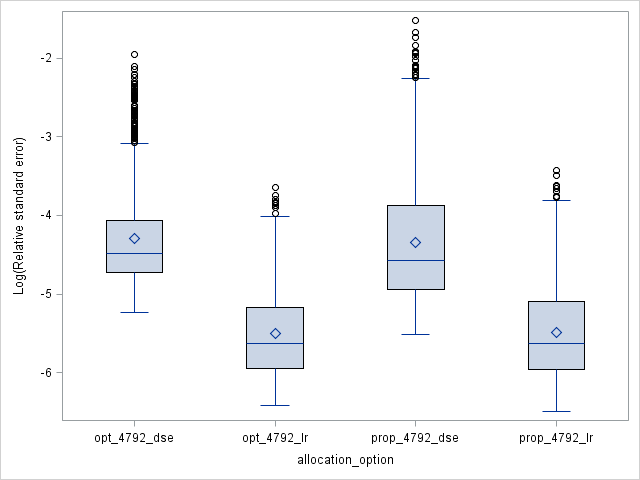
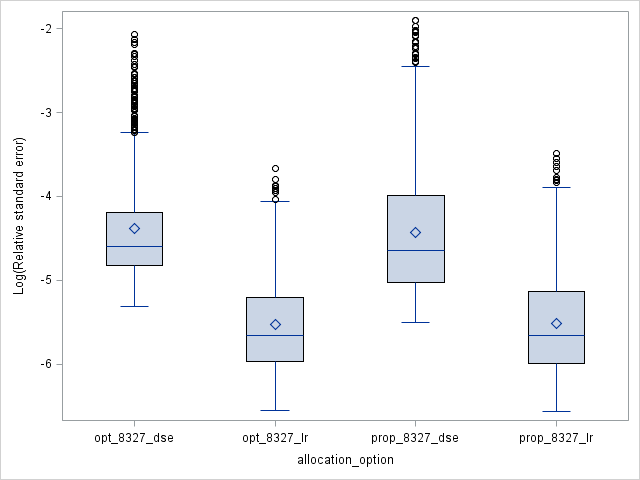
It is for the hybrid optimal-proportional allocation where there is almost consistently a reduction in variance with the less clustered design. Even in LA x HtC 1 and 2 strata the RSE are generally lower for the less clustered allocation. As with the optimal and proportional allocation methods, these differences become more pronounced as the HtC index increases.

Collectively, the results for the comparison of sampling fractions support the recommendation to propose a smaller SSU sampling fraction of 0.25.

*Comparison of estimation method with allocations*

Figure 13 permits a comparison of logistic regression estimation and dual system, ratio, synthetic estimators according to the allocation methods.

1. b)



*Figure 13: log(RSE) for all LA x HtC totals for optimal and proportional allocation. Box plots (a) PSU sampling fraction 0.0435, SSU sampling fraction 0.25; (b) PSU sampling fraction 0.0241, SSU sampling fraction 0.5. Logistic regression estimation (\_lr) and dual system, ratio, synthetic estimation (\_dse).*

As expected, the reduction in variance using the logistic regression estimation approach, compared to the dual system, ratio, synthetic estimators, is greater if the allocation strategy is proportional rather than optimal allocation. This finding holds with both sampling fractions. These findings support the expectation that there would be less requirement to allocate the sample disproportionately across the strata with a logistic regression model than with the dual system, ratio, synthetic approach.

*Comparison of constraints*

Results are displayed next to compare the optimal allocation strategy with and without maximum constraints.

Figure 14 and Figure 15 are log RSE for all LA x HtC strata estimates for optimal allocation option 1 (PSU sampling fraction 0.0241, SSU sampling fraction 0.5) and logistic regression estimation. Figure 16 and Figure 17 show the equivalent for option 2. These figures show that for logistic regression estimation, there are negligible differences in RSE between an allocation option with this maximum constraint, and one without.

1. b)

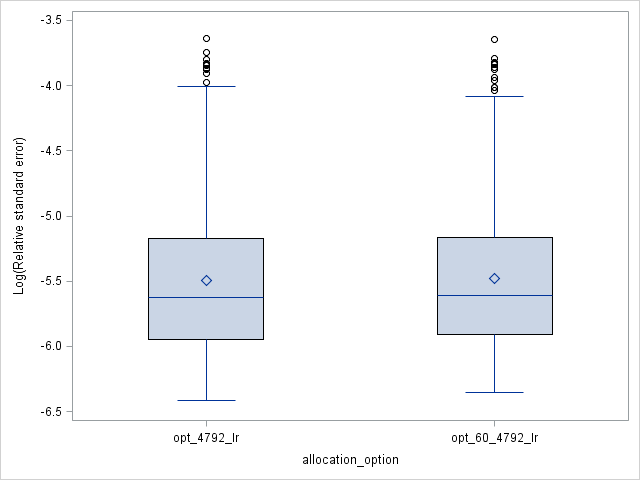
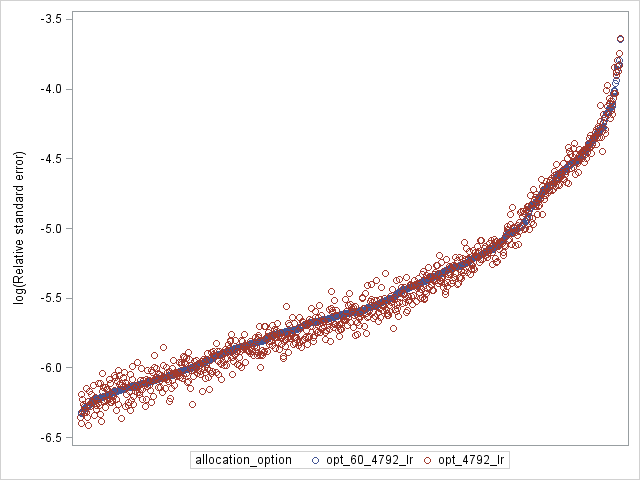


Figure 14: Log(RSE) for all LA x HtC totals for optimal allocation with PSU sampling fraction 0.0241, SSU sampling fraction 0.5, and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size for opt\_4792\_lr (unconstrained). For maximum 60 PSU per LA see opt\_60\_4792\_lr; (b) equivalent box plot.

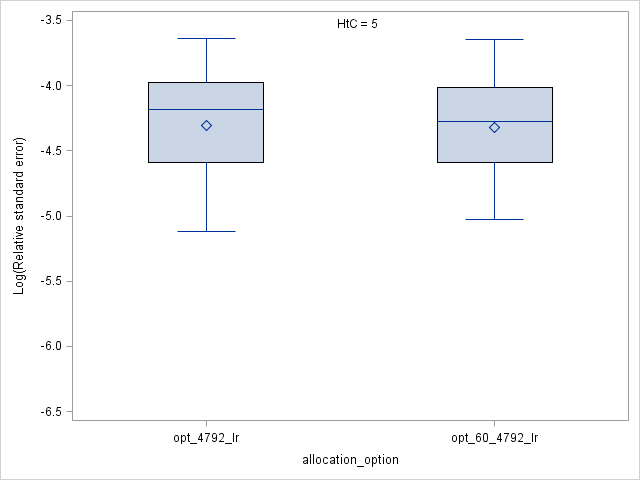
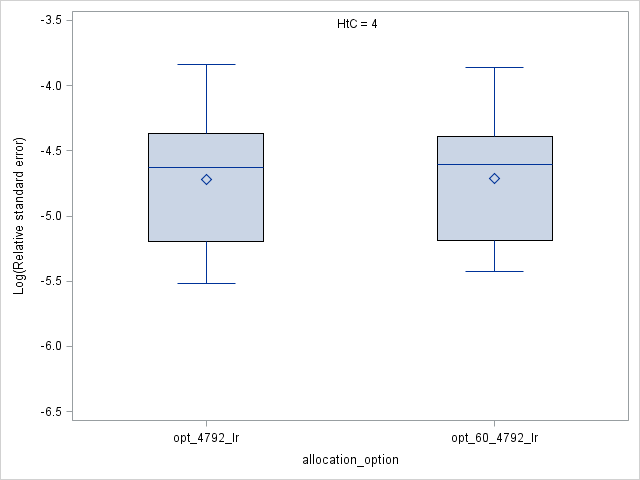
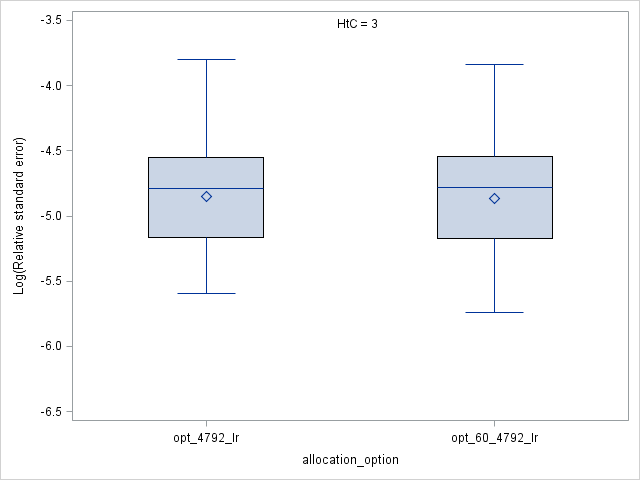
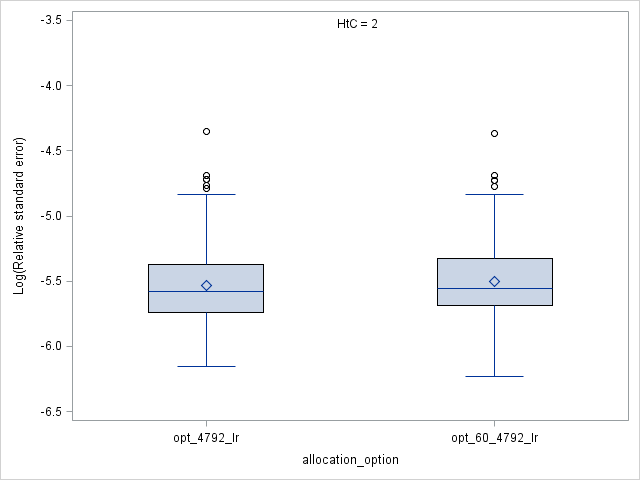
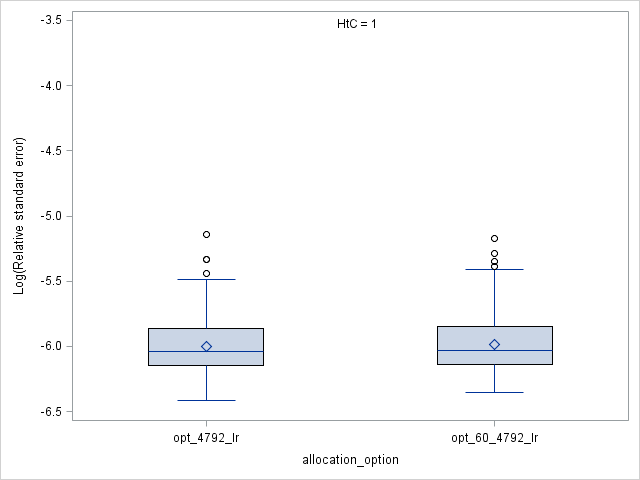


Figure 15: Box plots of log(RSE) for LA x HtC totals for optimal option 1 unconstrained (opt\_4792\_lr), and constrained (opt\_60\_4792\_lr) and displayed by HtC index category. Logistic regression model estimation.

1. b)

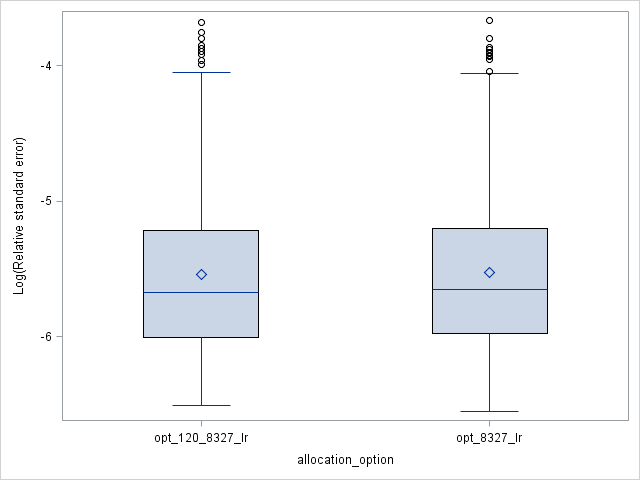
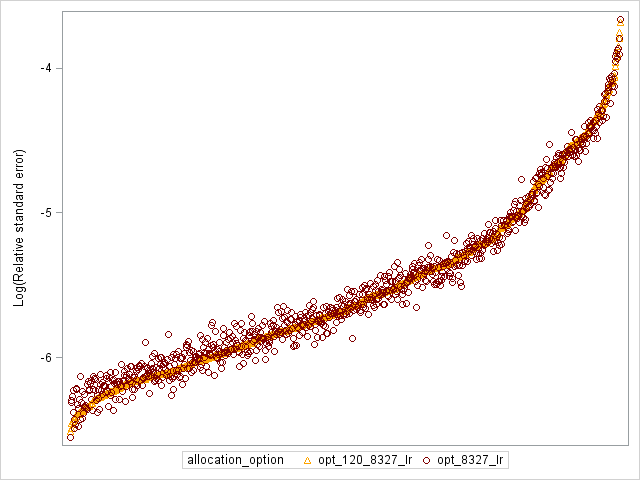


Figure 16: log(RSE) for all LA x HtC totals for optimal allocation with PSU sampling fraction 0.0435, SSU sampling fraction 0.25, and logistic regression model estimation. Scatter plot (a) ordered by RSE size for opt\_120\_8327\_lr for maximum 120 PSU per LA. Unconstrained is opt\_8327\_lr; (b) equivalent box plot.

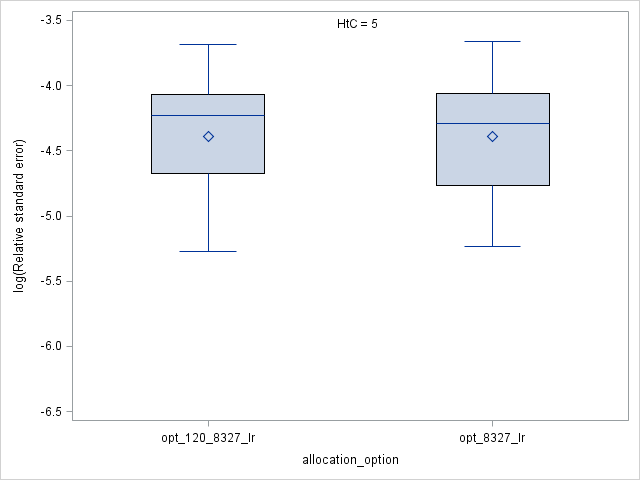
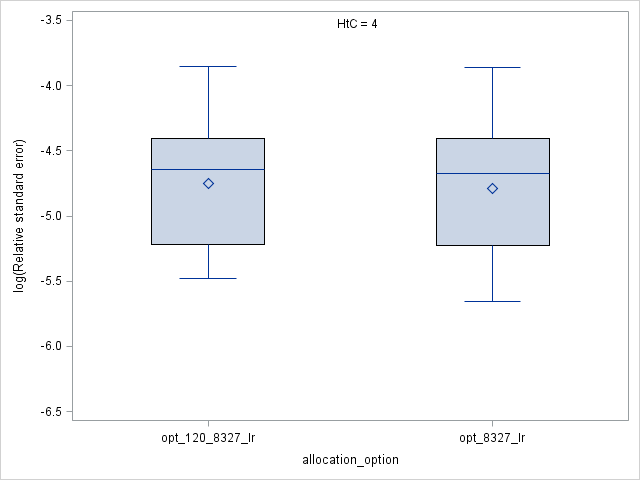
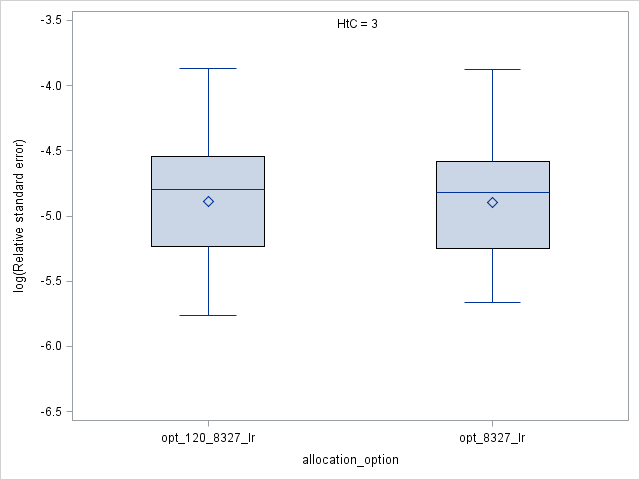
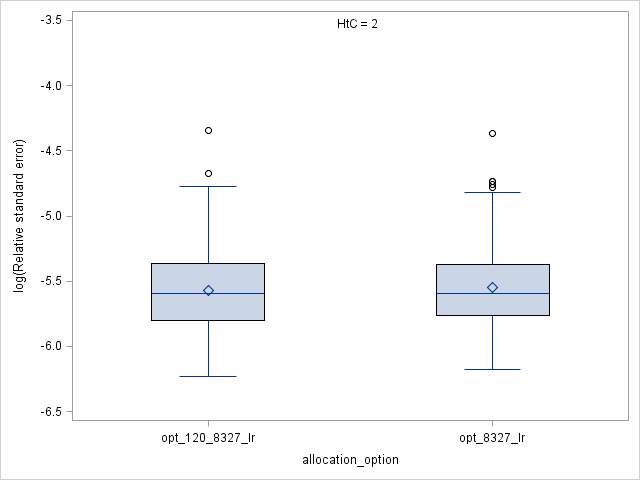
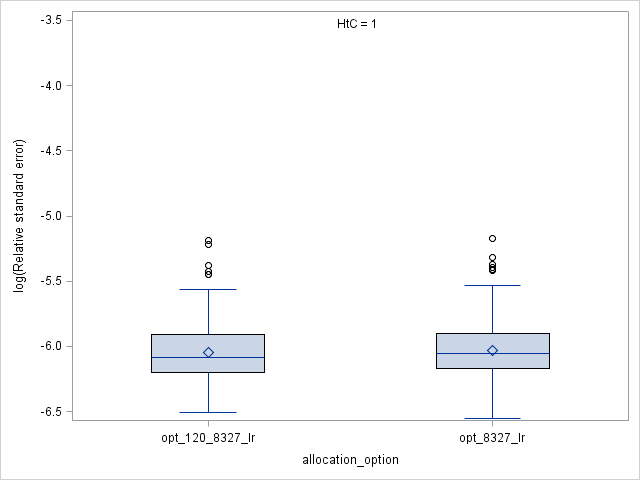


Figure 17: Box plots of log(RSE) for LA x HtC totals for optimal option 2 constrained (opt\_120\_8327\_lr), and unconstrained (opt\_8327\_lr) and displayed by HtC index category. Logistic regression model estimation.

Figure 18 and Figure 19 display log RSEs for all LA x HtC strata estimates for optimal allocation option 1 with PSU sampling fraction 0.0241, SSU sampling fraction 0.5, and dual system, ratio, synthetic estimators. Figure 20 and Figure 21 show the equivalent for option 2.

1. b)

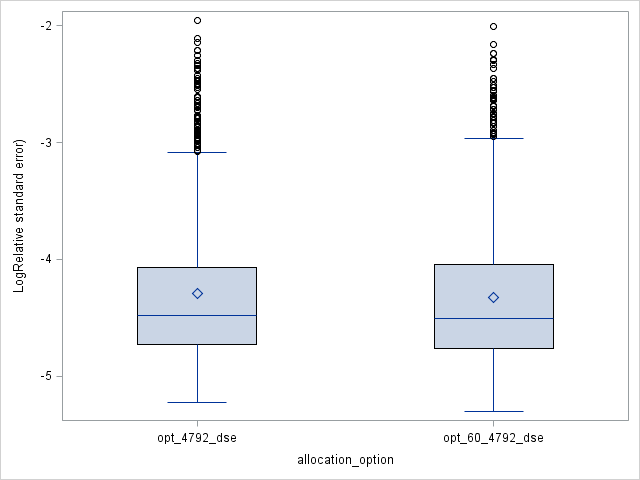
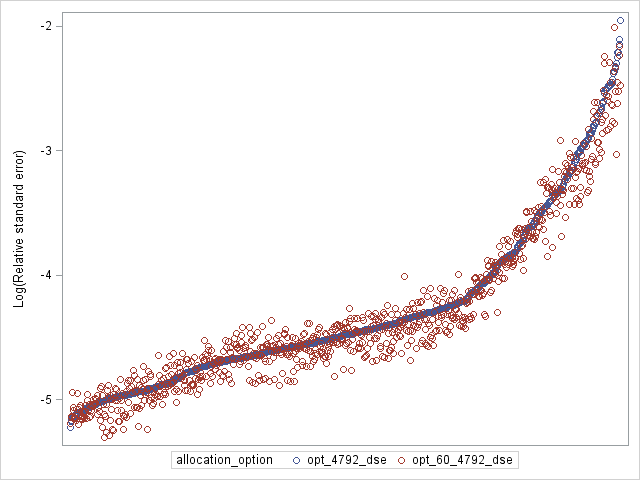


Figure 18: Log(RSE) for all LA x HtC totals for optimal allocation with PSU sampling fraction 0.0241, SSU sampling fraction 0.5, and dual system, ratio, synthetic estimation. Scatter plot (a) ordered by log(RSE) size for unconstrained opt\_4792\_dse. For maximum 60 PSU per LA see opt\_60\_4792\_dse; (b) equivalent box plot.

For estimation by dual system, ratio, synthetic estimators, there is evidence that there are smaller RSEs for the allocation option that constrains the maximum number of PSUs per LA. For the 2011 sampling fractions, there is supporting evidence of increased precision in LA x HtC 3, 4 and 5 categories, hence the decision to use a maximum constraint in the allocation strategy in 2011.

Collectively, the results for comparison of the maximum constraint for optimal allocation support a recommendation to propose the use of this constraint. The variances with the maximum constraint and logistic regression are no worse than those without, and they will benefit the operational process.

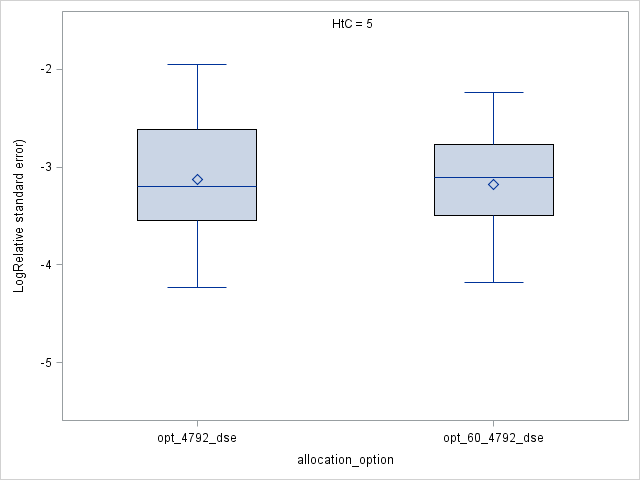
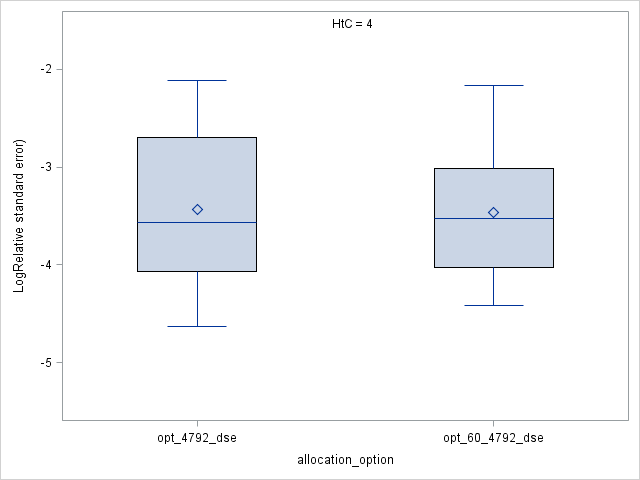
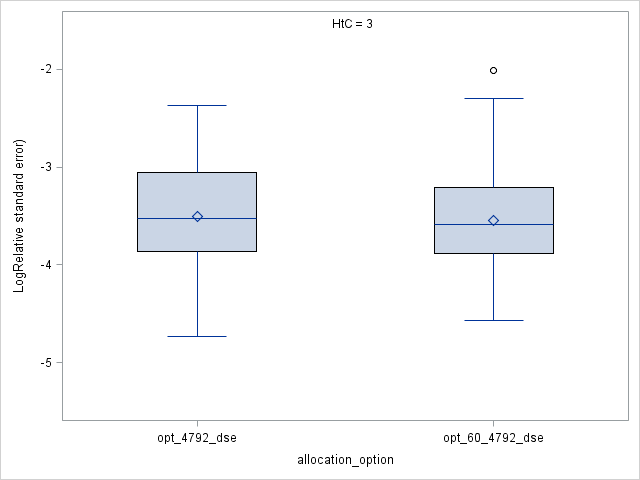
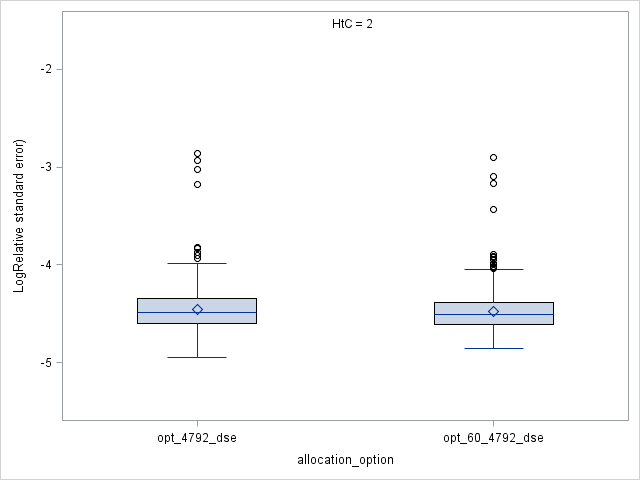
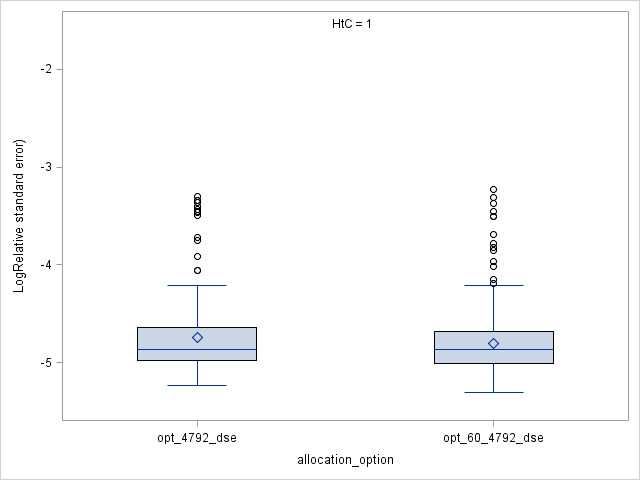


Figure 19: Box plots of log(RSE) for LA x HtC totals for optimal option 1 unconstrained (opt\_4792\_dse), and constrained (opt\_60\_4792\_dse) and displayed by HtC index category. Dual system, ratio, synthetic estimators.

1. b)

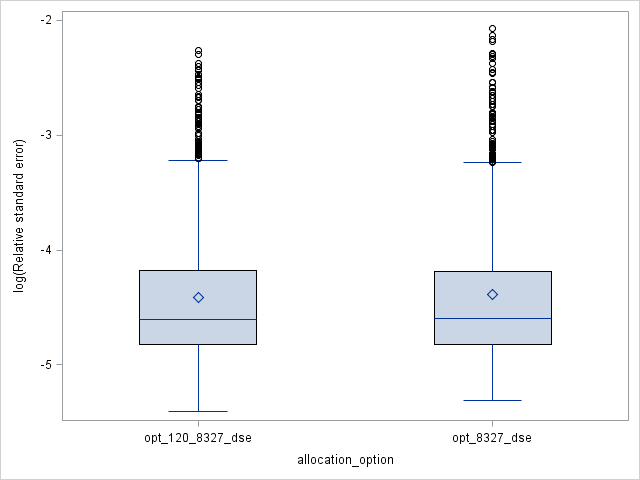
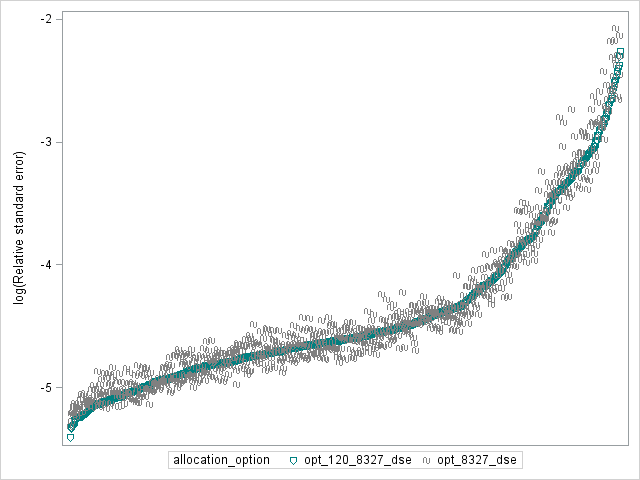


Figure 20: log(RSE) for all LA x HtC totals for optimal allocation with PSU sampling fraction 0.044, SSU sampling fraction 0.25, and dual system, ratio, synthetic estimation. Scatter plot (a) ordered by log(RSE) size for opt\_120\_8327\_dse for maximum 120 PSU per LA. Unconstrained is opt\_8327\_dse; (b) equivalent box plot.

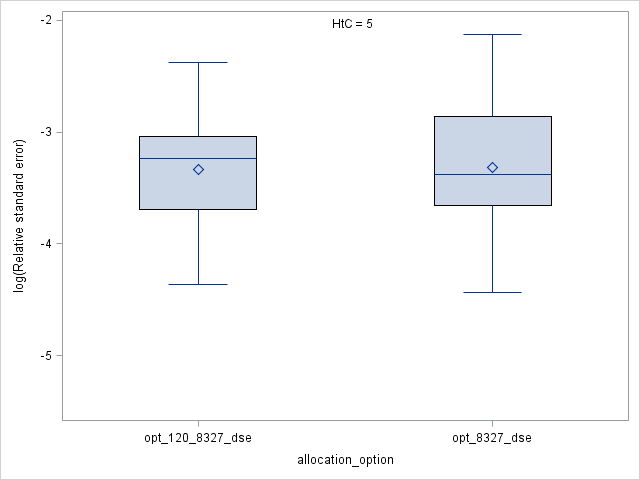
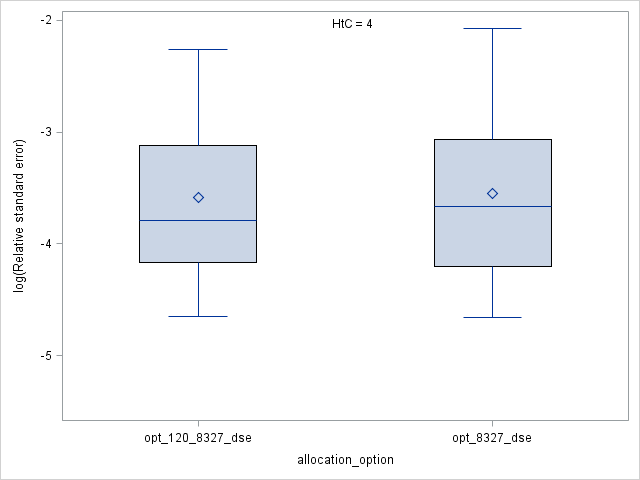
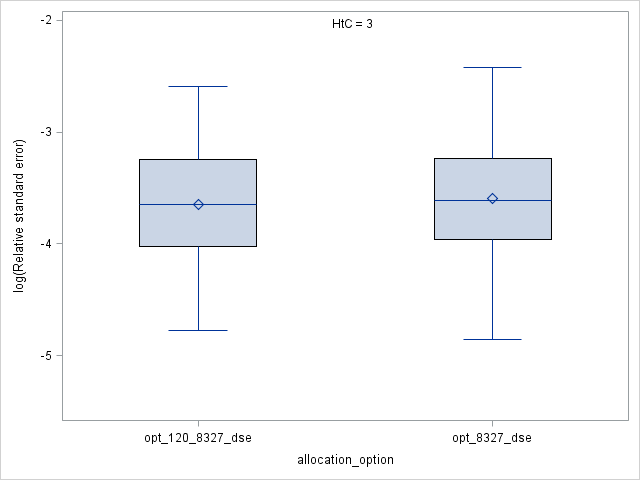
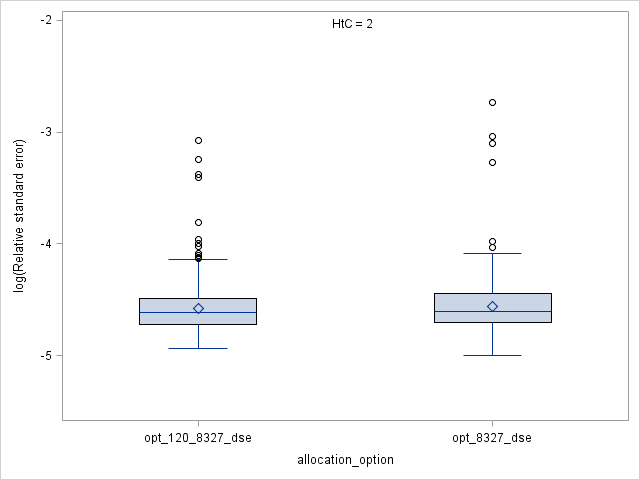
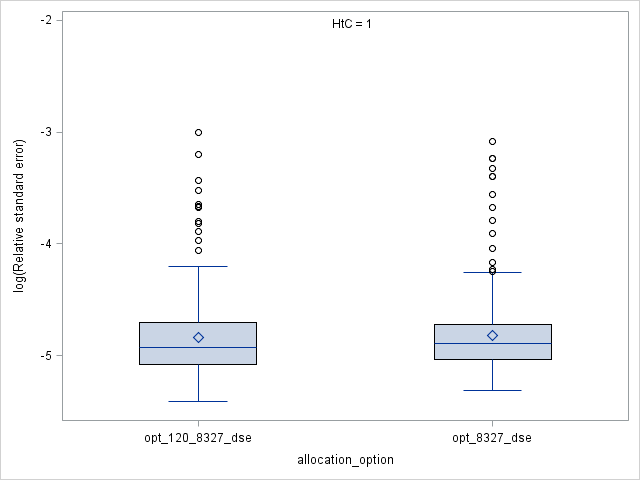


Figure 21: Box plots of log(RSE) for LA x HtC totals for optimal option 2 constrained (opt\_120\_8327\_dse), and unconstrained (opt\_8327\_dse) and displayed by HtC index category. Dual system, ratio, synthetic estimators.

* + 1. **Relative standard errors for LA totals (total of 348 points)**

*Comparison of allocation methods*

For LA total estimates, there are similar findings as for LA x HtC estimates, although relatively, the differences are of smaller magnitude since the domains are larger Corresponding figures shown below (note the change of *y-*axis scale compared to those for LA x HtC estimates).

1. b)

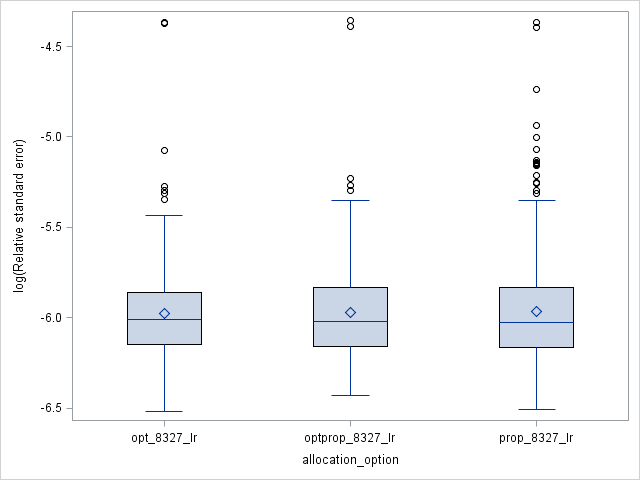
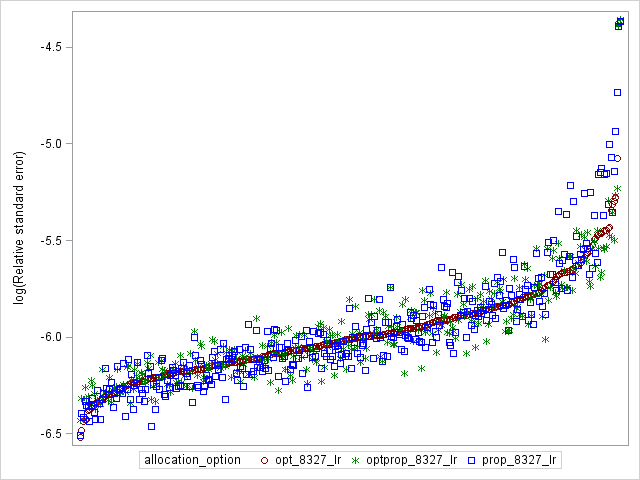


Figure 22: log(RSE) for all LA totals for optimal (opt\_8327\_lr), hybrid optimal/proportional allocation (optprop\_8327\_lr), and proportional (prop\_8327\_lr), and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size in optimal allocation with PSU sampling fraction 0.0435, SSU sampling fraction 0.25; (b) corresponding box plot.

*Comparison of sampling fractions*

*Optimal, proportional and optimal-proportional allocation*

1. b)

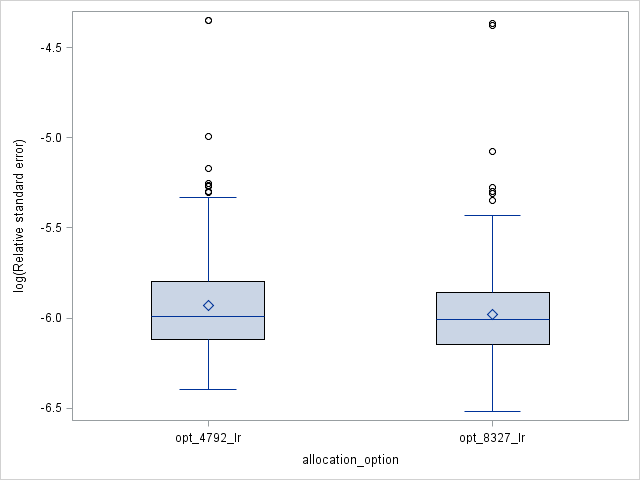
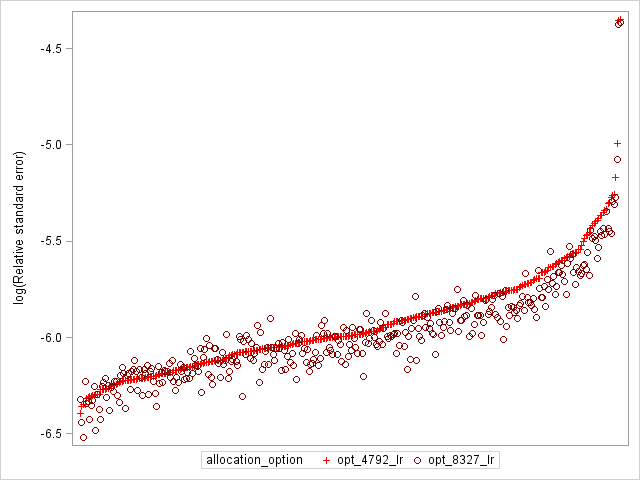


Figure 23: log(RSE) for all LA x HtC totals for optimal allocation options 1 and 2, and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (opt\_4792\_lr). PSU sampling fraction 0.0435, SSU sampling fraction 0.25 is opt\_8327\_lr; (b) corresponding box plot.

1. b)

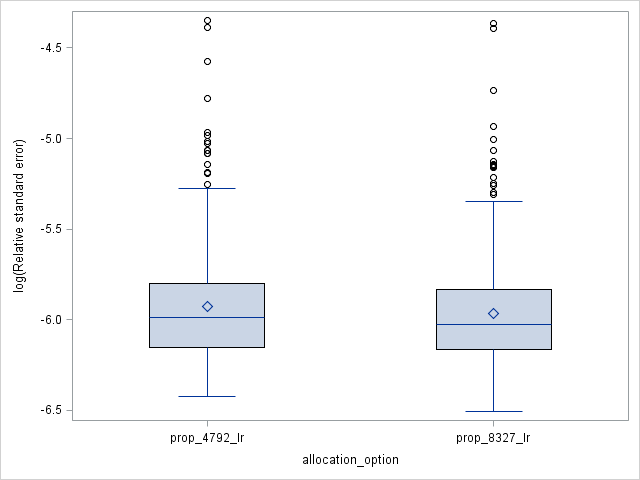
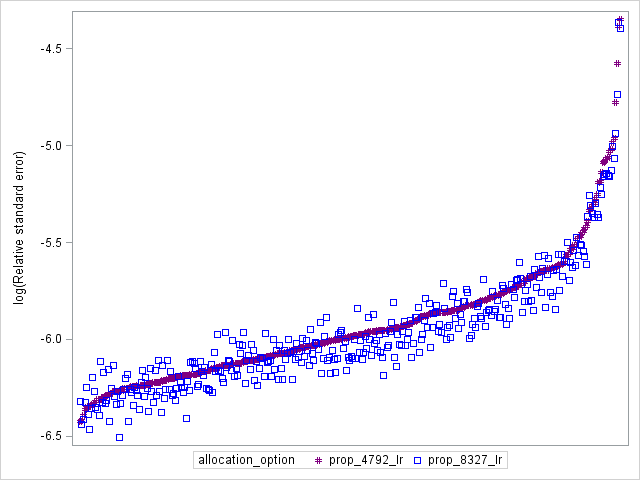


Figure 24: log(RSE) for all LA x HtC totals for proportional allocation options 3 and 4, and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (prop\_4792\_lr). PSU sampling fraction 0.0435, SSU sampling fraction 0.25 is prop\_8327\_lr; (b) corresponding box plot.

1. b)

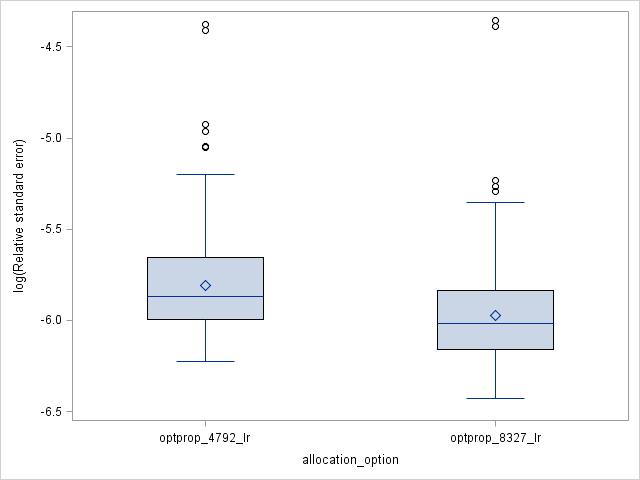
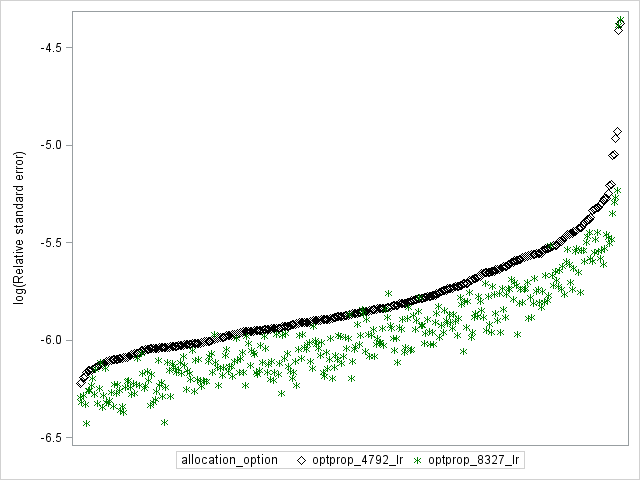


Figure 25: log(RSE) for all LA x HtC totals for optimal allocation options 5 and 6, and logistic regression model estimation. Scatter plot (a) ordered by log(RSE) size with PSU sampling fraction 0.0241, SSU sampling fraction 0.5 (optprop\_4792\_lr). PSU sampling fraction 0.0435, SSU sampling fraction 0.25 is optprop\_8327\_lr; (b) corresponding box plot.

## Discussion

Several sample allocation strategies for the 2021 census coverage estimation have been evaluated in the context of the 2021 Census coverage estimation strategy. All results obtained agree with what is expected from statistical theory. They give reliable quantification of performance of various allocation and sampling fraction strategies in the census coverage estimation context. Generally, optimal allocation has the lowest RSE, which is most prominent for the higher hard-to-count strata. However, also as expected, the differences are very small in the logistic regression estimation approach due to borrowing of strength. Largest variances are observed with proportional allocation of the CCS sample.

The study shows that the variances for the hybrid optimal-proportional allocation approach are similar to those for the optimal allocation. Therefore, we propose the use of this allocation approach, which will protect against over-optimising to 2011 patterns of census response, without having to be as conservative as to allocate the sample in a proportionate manner only.

The results here can be used to support the recommendation from earlier work and propose a smaller SSU sampling fraction of 0.25 (from 0.5 in 2011) to increase precision.

Another recommendation is to use the maximum allocation constraint of the 2011 strategy since the results indicate that it does not decrease precision with the logistic regression model approach. This protects over-allocating to some areas, which is particularly beneficial for fieldwork purposes.

The findings of this work should be considered in the context of other current work on modelling strategies that could be used to account for bias introduced in informative sampling that would be likely with an optimal-type allocation (Račinskij, 2020). Early findings demonstrate that there are strategies that would be able to cope with the informative sample selection in the 2021 CCS if this type of selection occurred.

Lastly, the findings of this study can be used to support the decision for having no flexible sample in the 2021 CCS sample. The rationale for a flexible sample in 2011 was that we cannot predict where the census response rates will be low. During the census operation, there may be pockets of low response in unexpected areas. The flexible sample approach in 2011 provided a boost sample of approximately 7000 addresses across 44 local authorities to include in existing workloads. The argument for no flexible sample in the 2021 CCS sample design is that the 2021 estimation approach will enable borrowing of strength across areas and improve precision of estimates. A flexible sample of a similar size to 2011 will not be large enough to improve the estimates beyond the precision that can be gained through the borrowing of strength. Furthermore, the hybrid approach spreads the sample more evenly across the areas compared to optimal allocation by LA x HtC.

**References**

Abbott O. (2008b) CCS Sample Size and Allocation. Internal ONS document.

Abbott O., and J. Brown (2008) CCS Design Strategy for the 2011 Census. Internal ONS document.

Brown, J., Abbott, O., Smith, P.A. (2011) Design of the 2001 and 2011 Census Coverage Surveys for England and Wales. J.R. Statist. Soc. A, 174; 881-906.

Castaldo, A., Nikolakis, D. (2018) Assessing the use of an address-based design for the 2021 Census Coverage Survey. Internal ONS document.

Lohr, S. L. (1999) *Sampling: Design and Analysis*. Pacific Grove: Duxbury.

Office for National Statistics (2010) 2011 Census Coverage assessment: CCS sample sizes and Estimation Areas for local authorities – autumn 2010. Available at

<http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-consultations/uag/census-advisory-groups/statistical-development/census-coverage-assessment---ccs-sample-sizes.pdf>

Office for National Statistics (2017a) 2017 Census Test Report. Available from <https://www.ons.gov.uk/census/censustransformationprogramme/2017censustest/2017censustestreport> [Accessed 8 August 2019].

Office for National Statistics (2017b) 2021 Census topic research. Available from <https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/2021censustopicresearchdecember2017> [Accessed 8 August 2019].

Office for National Statistics (2018). 2021 Census Coverage Survey Design Strategy – V4.0. Internal ONS document.

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

Račinskij, V. (2020). Dealing with informative sampling in the coverage estimation of the 2021 Census of England & Wales. Internal ONS document.

**APPENDICES**

#### The coverage assessment strategy (Taken from Viktor’s Coverage Estimation Strategy for 2021, v1.7):

The main goal of the 2021 CCE is to provide the high-quality coverage error-corrected population totals and the variance of the population totals at the national and certain subnational levels. The preliminary quality targets for the population size estimates are as follows: the relative 95% confidence interval for the national population size estimate less or equal to ±0.2%, the absolute relative bias at the national level less than 0.5%; the relative 95% confidence intervals for the LA totals less or equal to ±3.0%.

#### Distribution of standard deviations, Sh, for optimal allocation

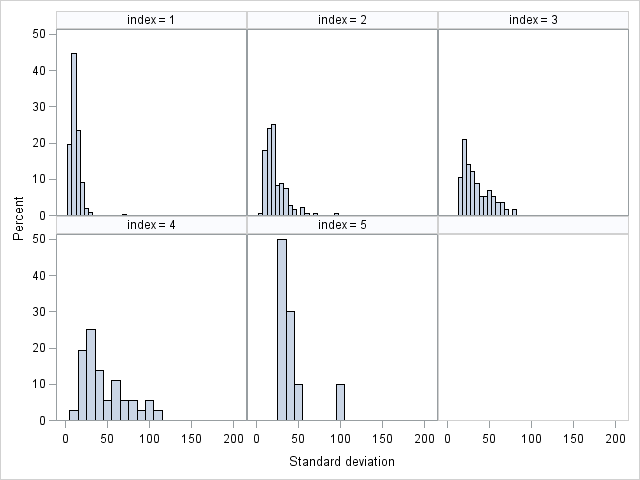


Figure B1: Distribution of standard deviations, Sh, for optimal allocation

#### Comparison of sample sizes across allocation options

Table C1: Comparison of the allocation sizes in optimal allocation option 1 and proportional allocation option 3.

|  |  |
| --- | --- |
| Difference in allocation size (number of OAs)  (optimal minus proportional) | Frequency of difference |
| -8 | 3 |
| -7 | 2 |
| -6 | 7 |
| -5 | 10 |
| -4 | 31 |
| -3 | 96 |
| -2 | 154 |
| -1 | 196 |
| 0 | 146 |
| 1 | 65 |
| 2 | 28 |
| 3 | 21 |
| 4 | 19 |
| 5 | 17 |
| 6 | 6 |
| 7 | 7 |
| 8 | 3 |
| 9 | 6 |
| 10 | 2 |
| 11 | 3 |
| 12 | 4 |
| 13 | 2 |
| 14 | 4 |
| 15 | 3 |
| 17 | 1 |
| 18 | 1 |
| 19 | 3 |
| 21 | 1 |
| 22 | 1 |
| 23 | 1 |
| 24 | 1 |
| 27 | 1 |
| 29 | 2 |
| 43 | 1 |

Table C2: Comparison of the allocation sizes in optimal allocation option 1 and optimal-proportional allocation option 5.

|  |  |
| --- | --- |
| Difference in allocation size (number of OAs)  (optimal minus optimal-proportional) | Frequency of difference |
| -21 | 1 |
| -14 | 1 |
| -13 | 1 |
| -12 | 1 |
| -10 | 1 |
| -9 | 1 |
| -8 | 5 |
| -7 | 7 |
| -6 | 4 |
| -5 | 1 |
| -4 | 10 |
| -3 | 9 |
| -2 | 34 |
| -1 | 188 |
| 0 | 396 |
| 1 | 90 |
| 2 | 37 |
| 3 | 21 |
| 4 | 9 |
| 5 | 5 |
| 6 | 5 |
| 7 | 6 |
| 8 | 4 |
| 9 | 5 |
| 10 | 1 |
| 12 | 2 |
| 13 | 2 |
| 23 | 1 |

#### Extra figures from simulation study

D1: Comparison of allocation methods with 2011 sampling fractions

1. b)

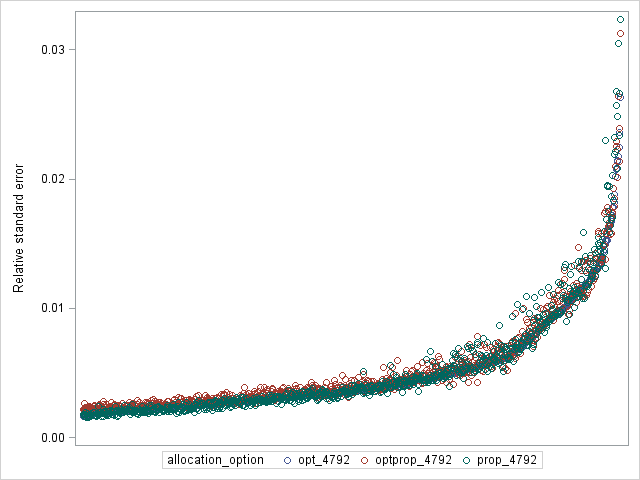
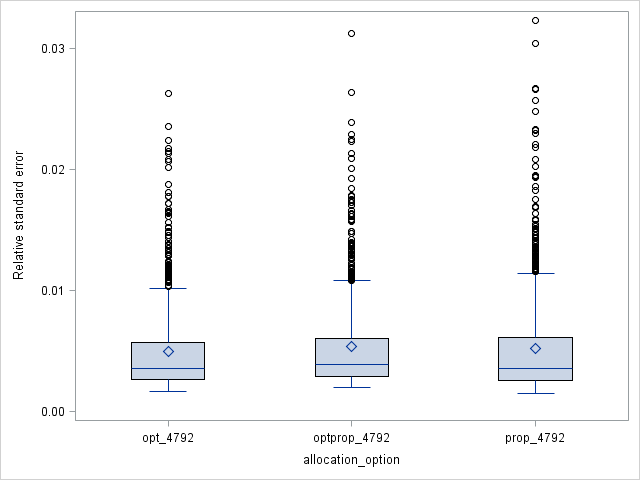
 

Figure D1: RSE for all LA / HtC totals for optimal (opt\_4792), proportional (prop\_4792), and hybrid optimal/proportional allocation (optprop\_4792), and logistic regression model estimation. Scatter plot (a) ordered by RSE size in optimal allocation with PSU sampling fraction 0.0241, SSU sampling fraction 0.5; (b) corresponding box plot.

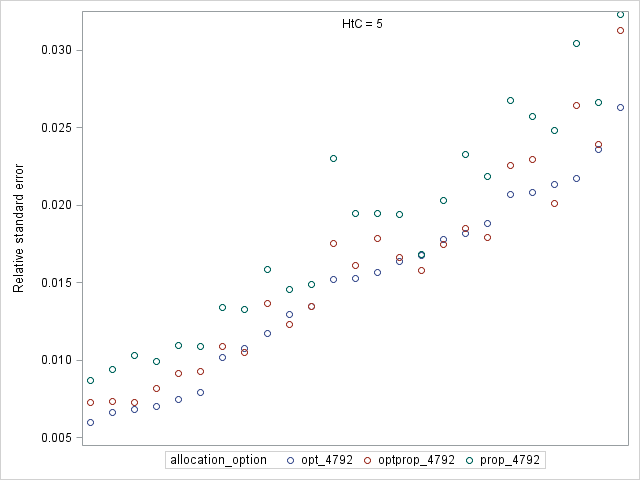
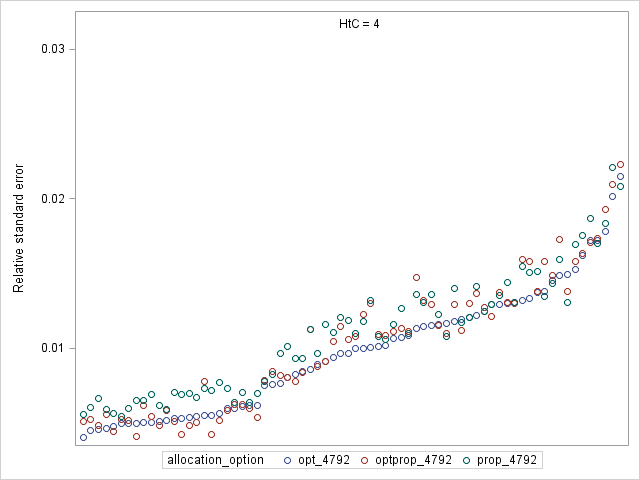
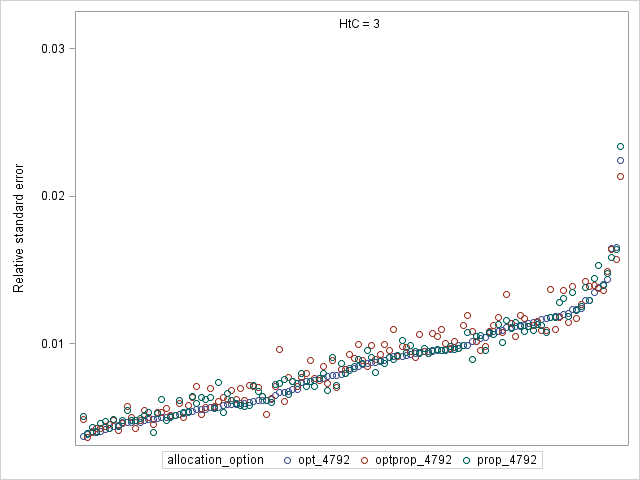
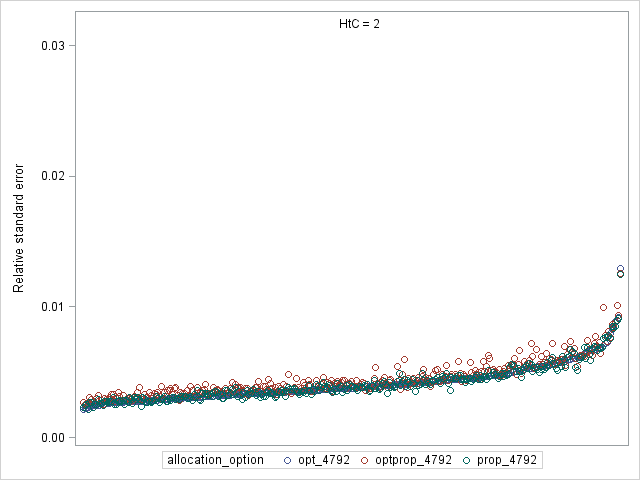
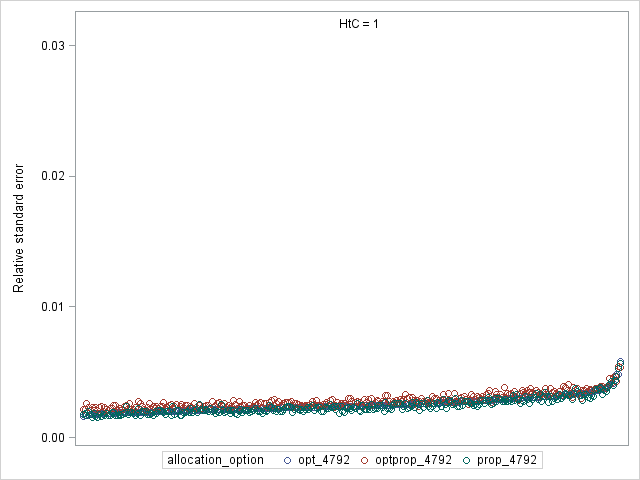


Figure D2: Scatter plots of RSE for LA / HtC totals for optimal (opt\_4792), proportional (prop\_4792), and hybrid optimal/proportional allocation (optprop\_4792), ordered by RSE size (optimal allocation) and displayed by HtC index category. Logistic regression model estimation.

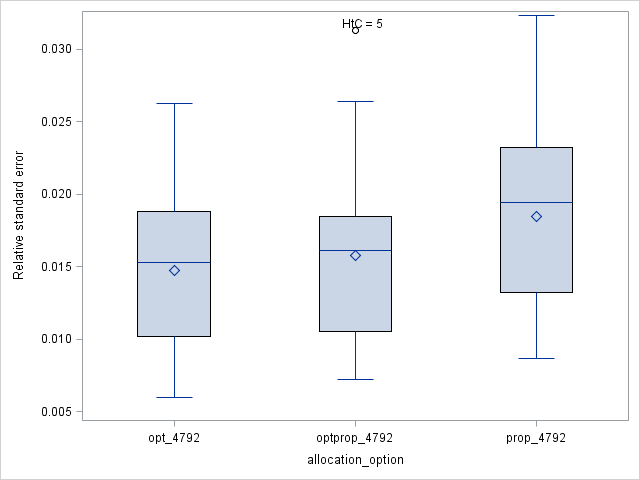
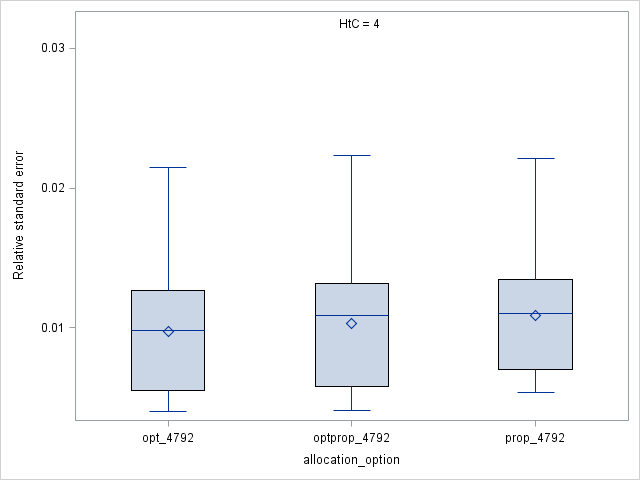
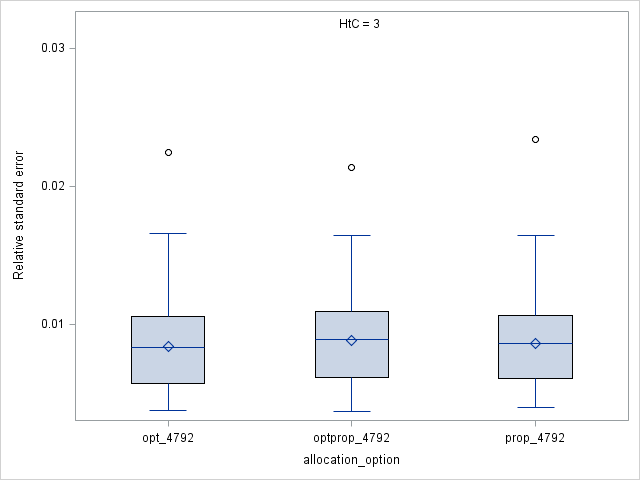
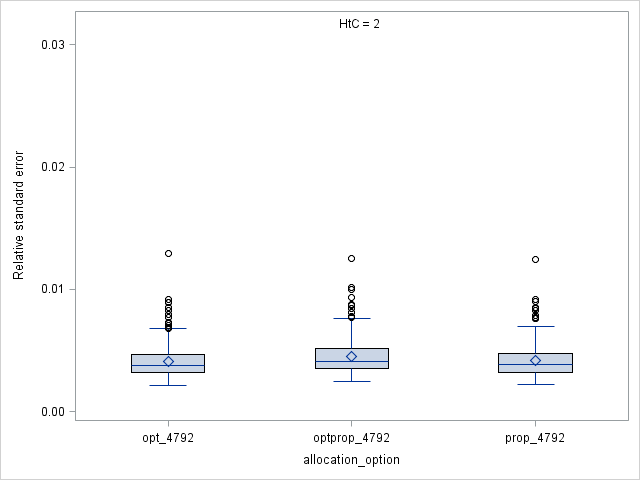
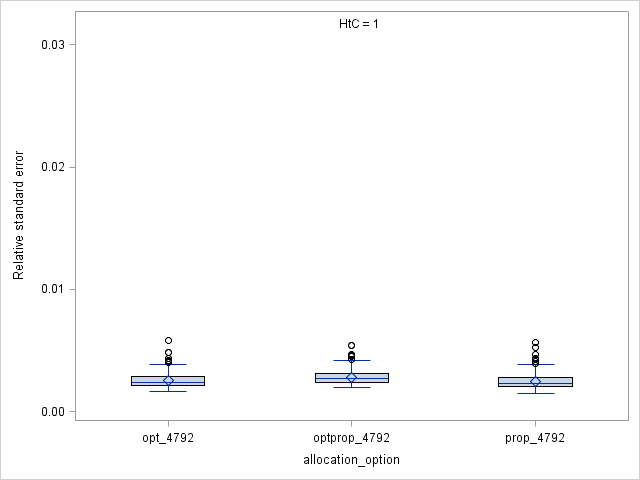


Figure D3: Box plots of RSE for LA / HtC totals for optimal (opt\_4792), proportional (prop\_4792), and hybrid optimal/proportional allocation (optprop\_4792) and displayed by HtC index category. Logistic regression model estimation.

D2: Comparison of sampling fractions with respect to estimation method

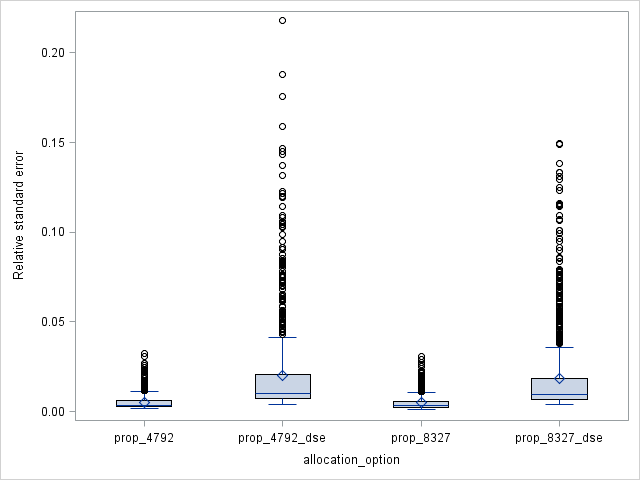
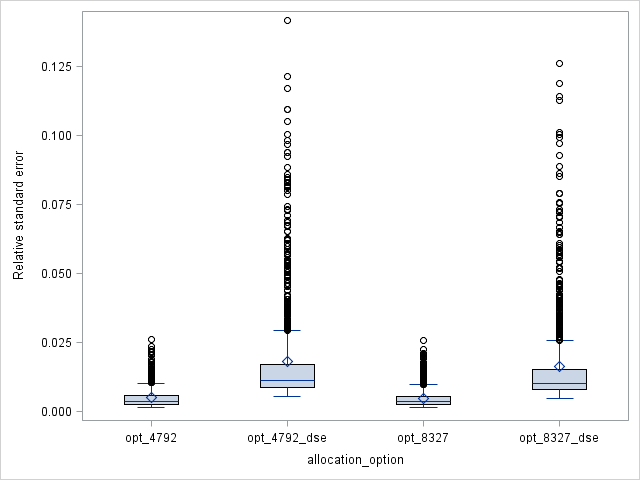
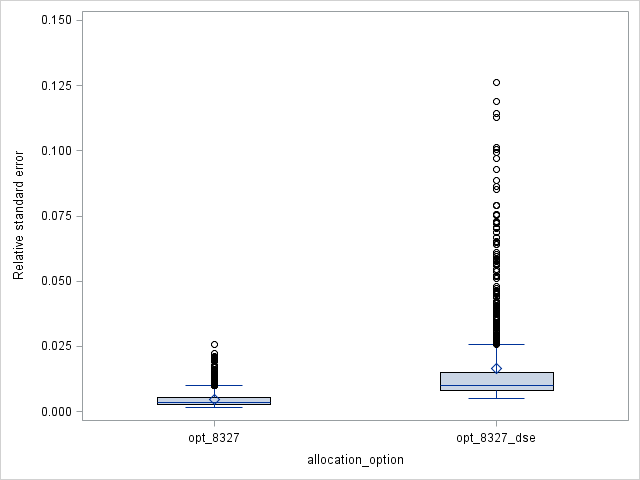
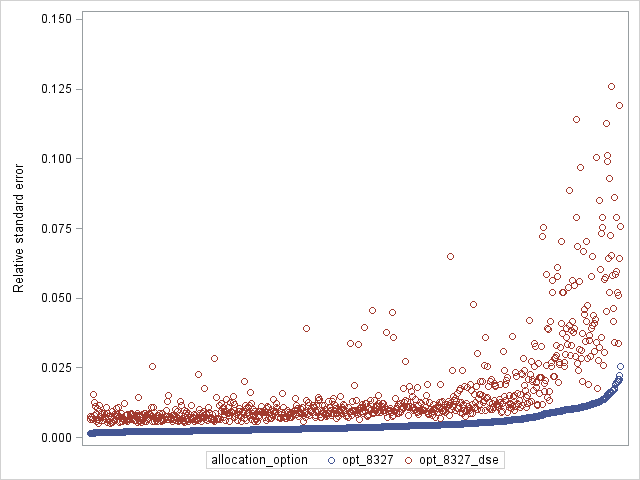


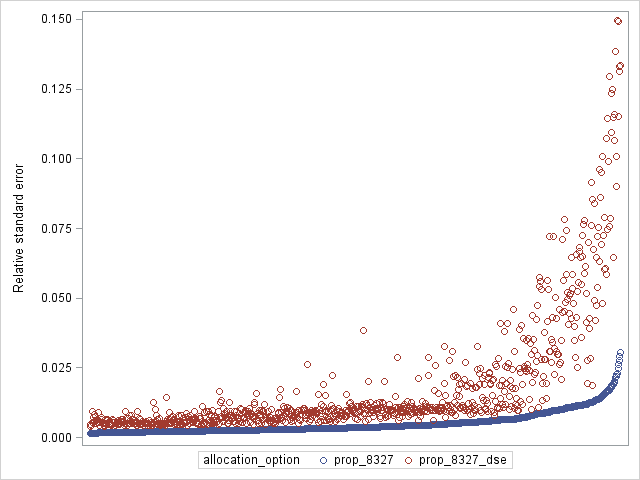
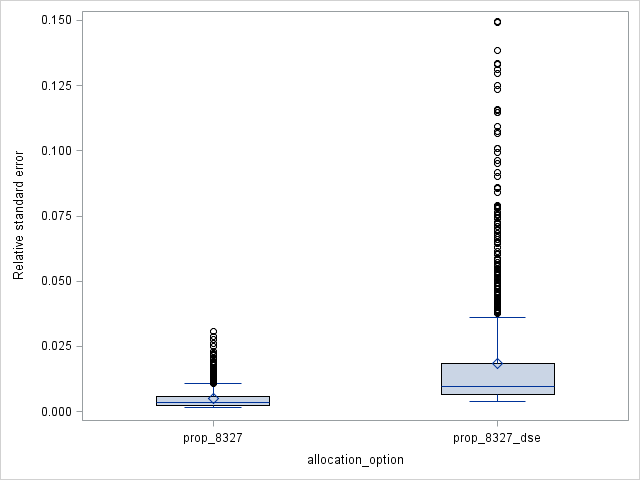
Figure D4: Box plots of RSE for LA / HtC totals for optimal and proportional allocation comparing sample fractions and estimation method.

D3: Scatter and box plots to compare estimation method for allocations: optimal, proportional and optimal-proportional. Sampling fractions PSU=0.0435, SSU=0.25, and PSU=0.0241, SSU=0.5.

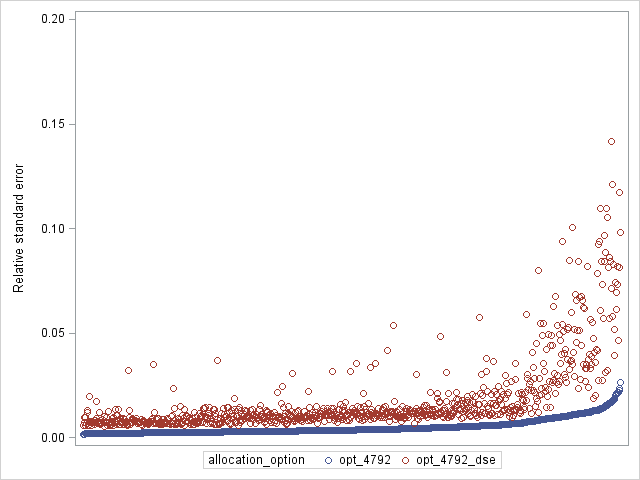
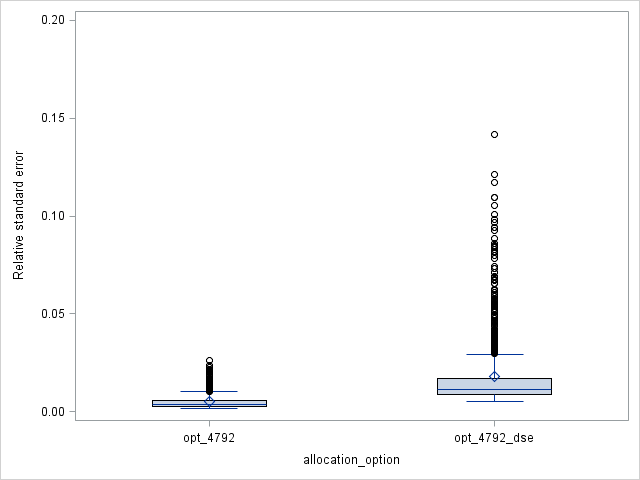
Optimal, 8327 OA, 25% postcodes



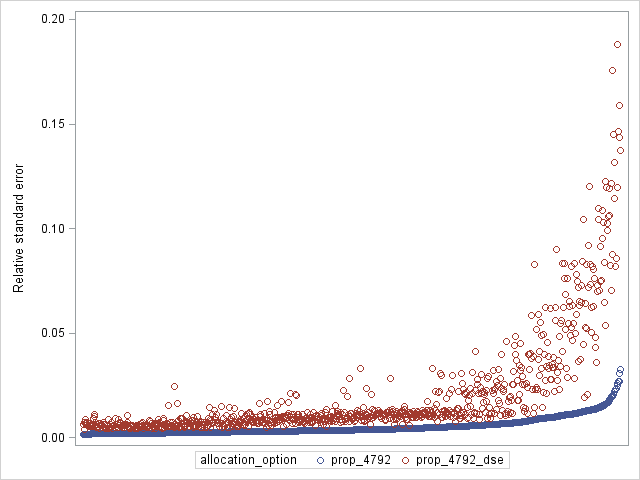
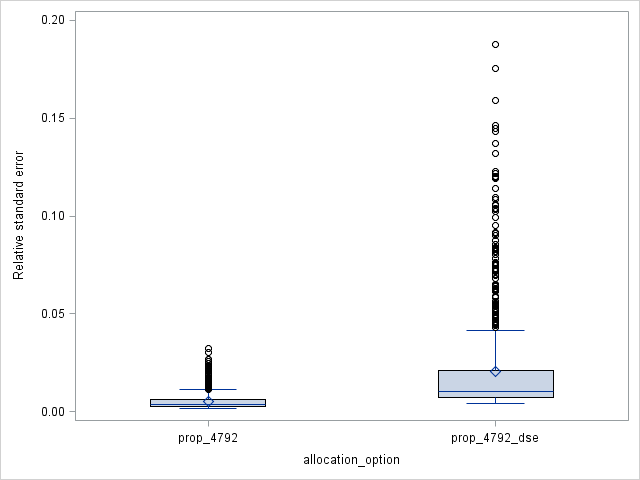
Proportional, 8327 OA, 25% postcodes

Optimal, 4792 OA, 50% postcodes

Proportional, 4792 OA, 50% postcodes

#### Early simulation work

E1: Additional methods

Simulation methods used in the early work are as those described in the main paper. However, as well as proportional and optimal allocation, a further two allocation options were considered. The simulated data used to generate the coverage probabilities for the allocation is also used to generate the response probabilities in the census data. Therefore, the data is perfectly (or “super”) optimal for the census data. In some sense, this is good because we can be sure of what the perfect situation looks like; it gives us a benchmark. However, the true variance of response probabilities is unknown, and we need to assess the effect of allocating the sample in what is likely to be a less optimal way. This will help to understand how conservative we may need to be with allocating the sample using 2011 Census data, should the response patterns be unlike the realised response rates in 2021.

Briefly, we did this by reversing / switching the standard deviations of coverage patterns in the strata. In general, the variability in response to census is smallest and least variable in HtC 1 strata, with both size and variability of the variance increasing as the HtC index increases (see figure B1 below). It is these standard deviations that largely drive the allocation sizes to each stratum in the optimal allocation approach. Therefore, to observe the kind of effect of allocating the sample in a less optimal way, we have used a resampling method to draw response probabilities for those in HtC 1 strata from individuals in HtC 5 strata, and vice versa, and then draw response probabilities for those in HtC 2 strata from individuals in HtC 4 strata, and vice-versa (HtC 3 remain the same). This has the effect of switching the distributions of variability in response in HtC 1 strata with those from HtC 5, and similarly for HtC 2 with HtC 4. A second option considered was a less exaggerated form, where it was just HtC 1 and 5 that were altered.

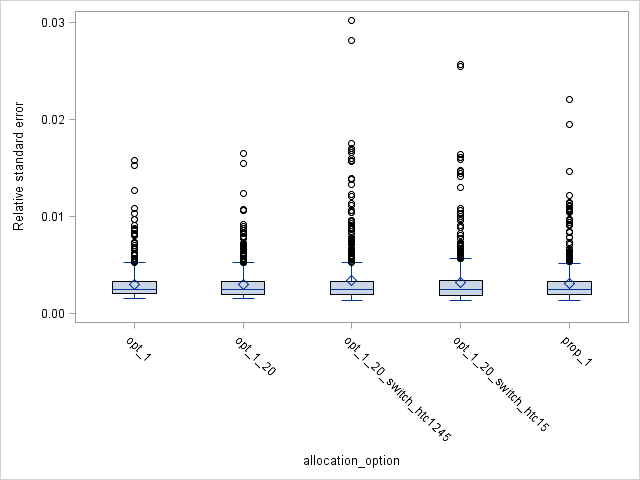
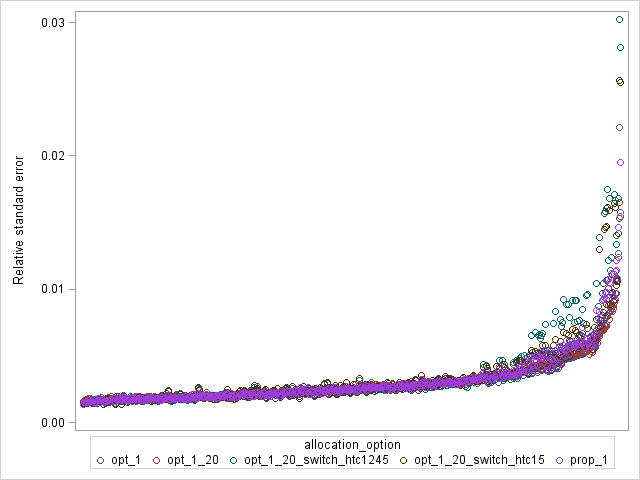
Five options considered in early work:

1. Optimal allocation based on simulated 2011 Census coverage patterns, min=1 PSU per LA / HtC stratum
2. Optimal allocation based on simulated 2011 Census coverage patterns, min=1 PSU per LA / HtC stratum, max=20 LSOA per LA
3. Proportional allocation, min=1 min=1 PSU per LA / HtC stratum, max=20 LSOA per LA
4. Optimal allocation with switched standard deviations for HtC 1 with HtC 5, and HtC 2 with HtC 4.
5. Optimal allocation with switched standard deviations for sample size for HtC 1 with HtC 5.

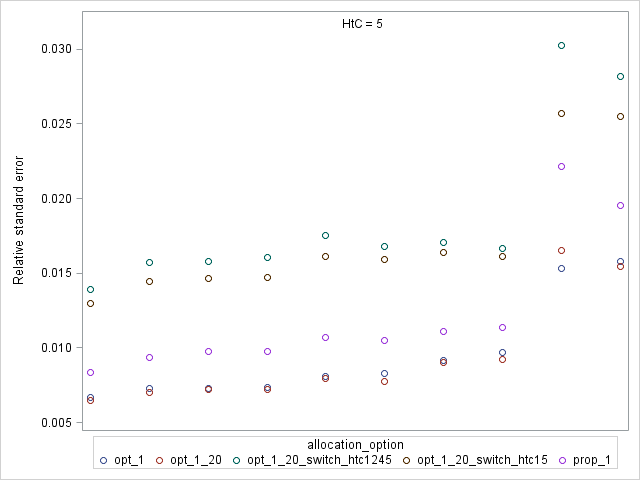
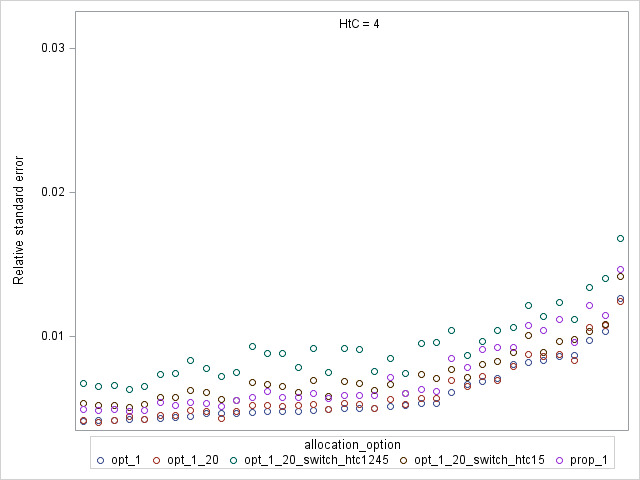
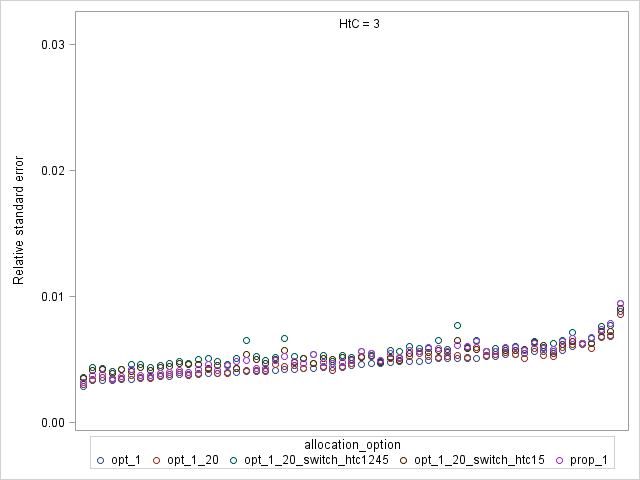
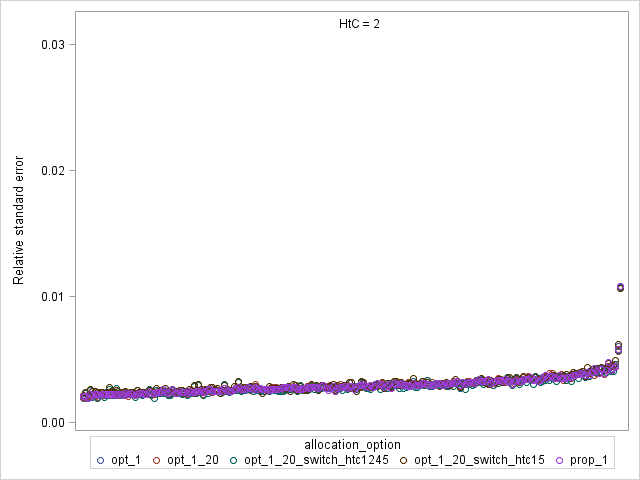
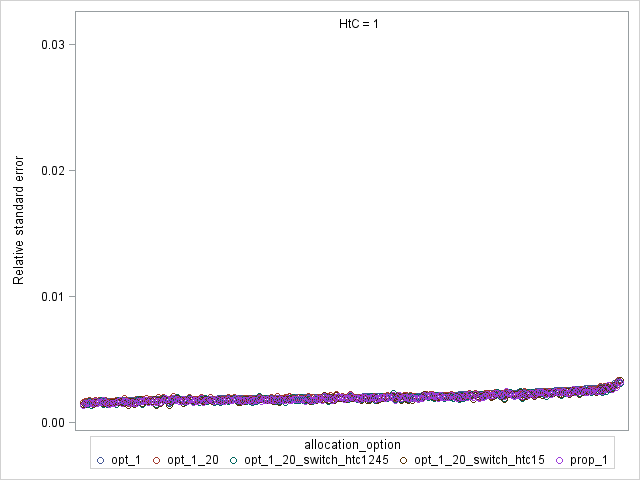
The findings from this early work showed that an optimal allocation approach that is not optimised for realised patterns can be worse in terms of variability compared to proportional allocation.

E2: Figures from early work

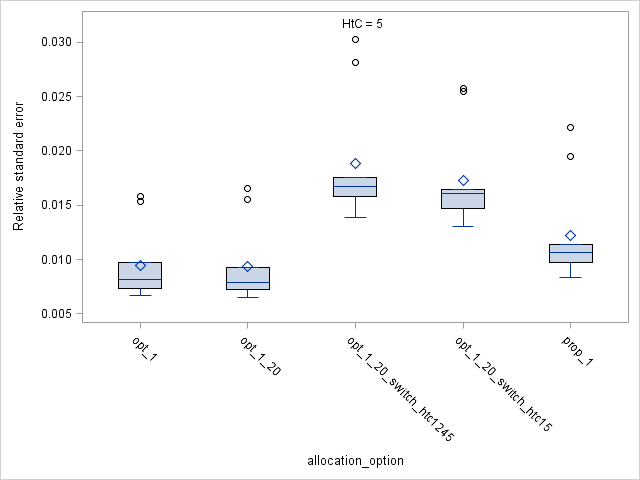
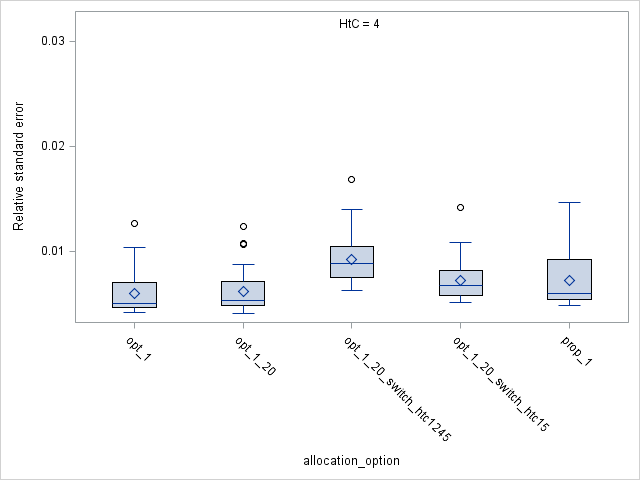
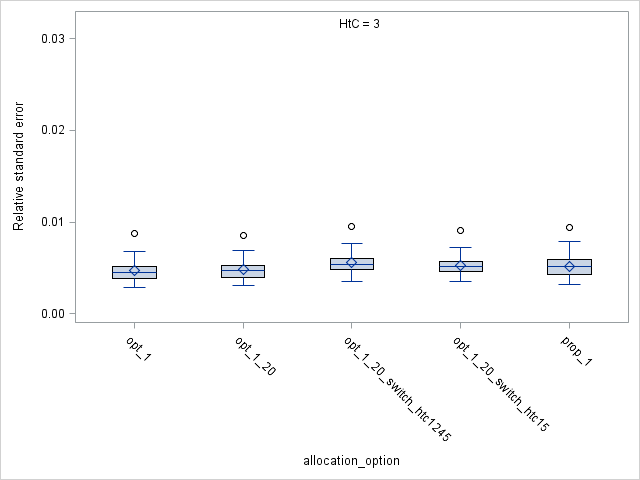
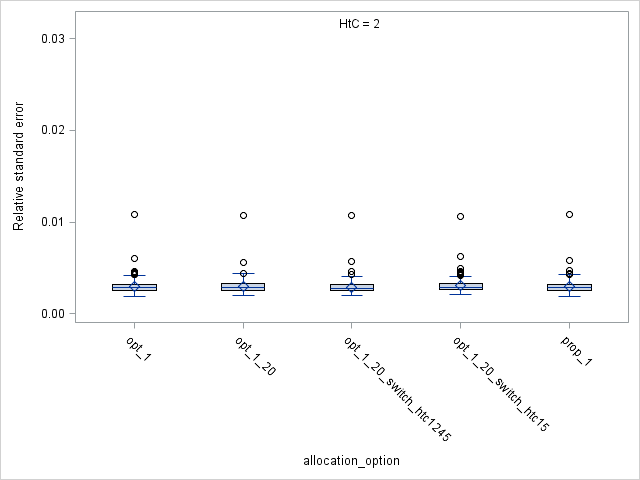
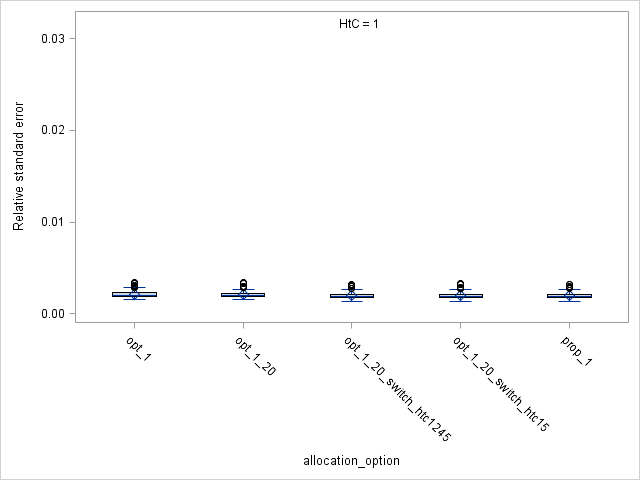
1. b)



*Figure E1: RSE for all LA / HtC totals for options 1 to 5 assuming CCS non-response and fitted with the logistic regression model. Scatter plot (a) ordered by RSE size in option 1.*

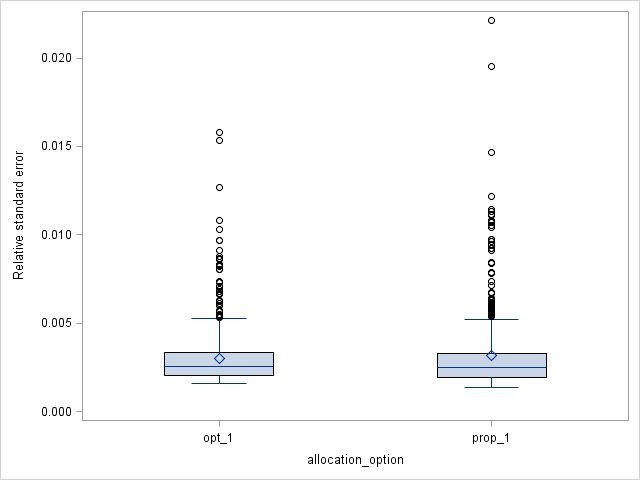
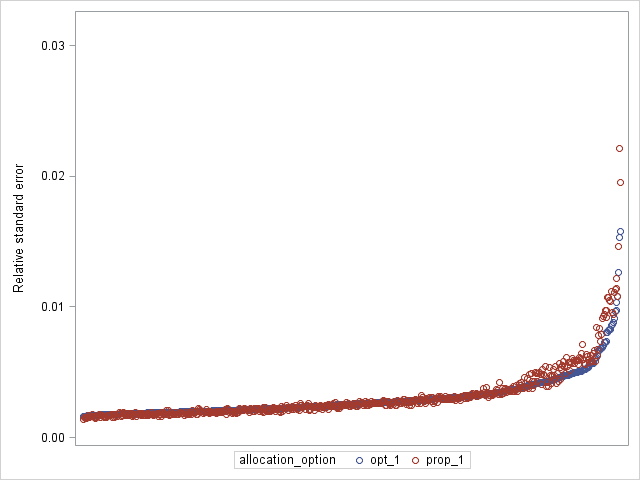


*Figure E2: Scatter plots of RSE for LA / HtC totals for options 1 to 5 assuming CCS non-response, ordered by RSE size (option 1) and displayed by HtC index category. All fitted with the logistic regression model.*

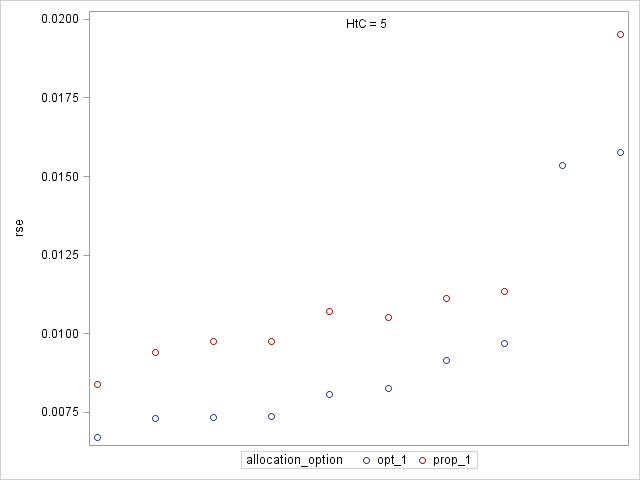
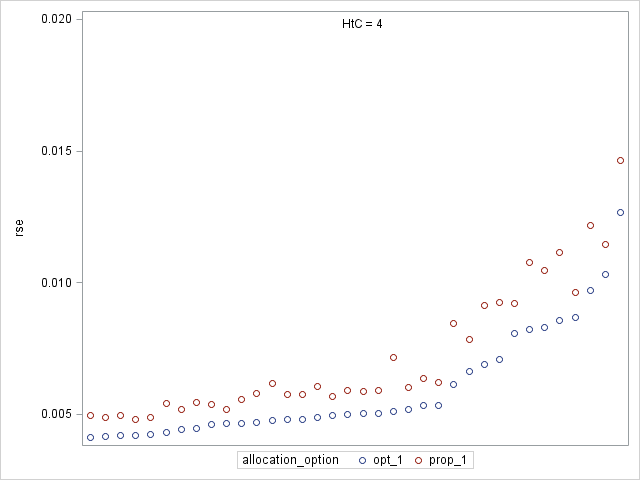
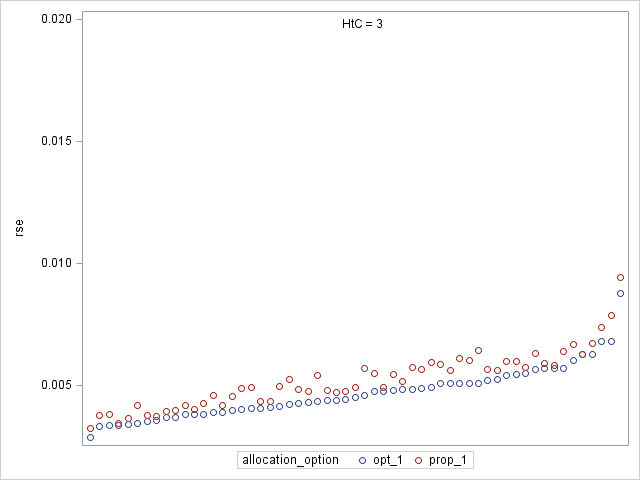
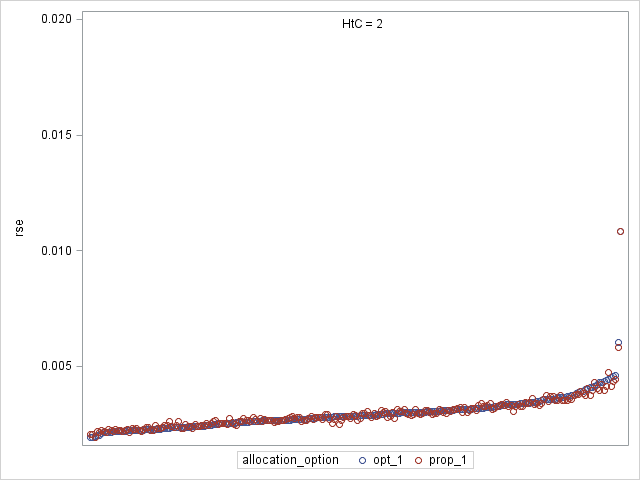
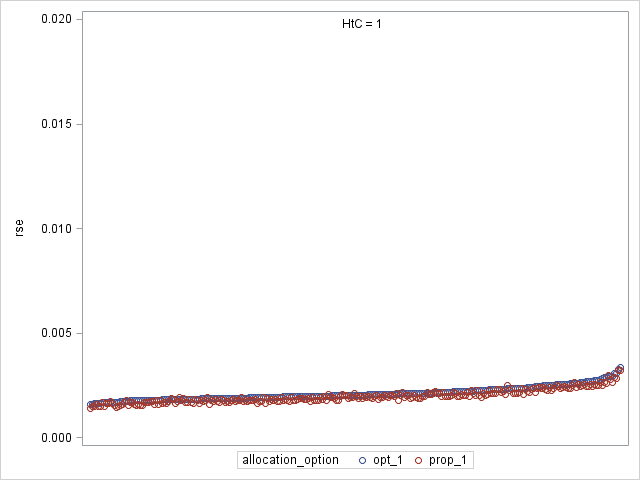


*Figure E3: Box plots of RSE for LA / HtC totals for options 1 to 5 assuming CCS non-response and displayed by HtC index category. All fitted with the logistic regression model.*

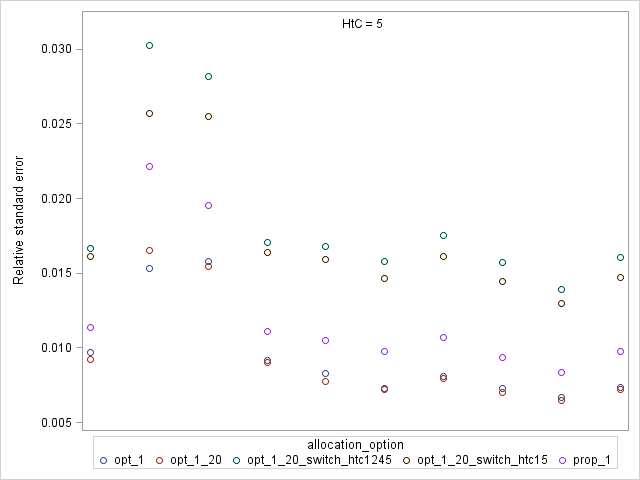
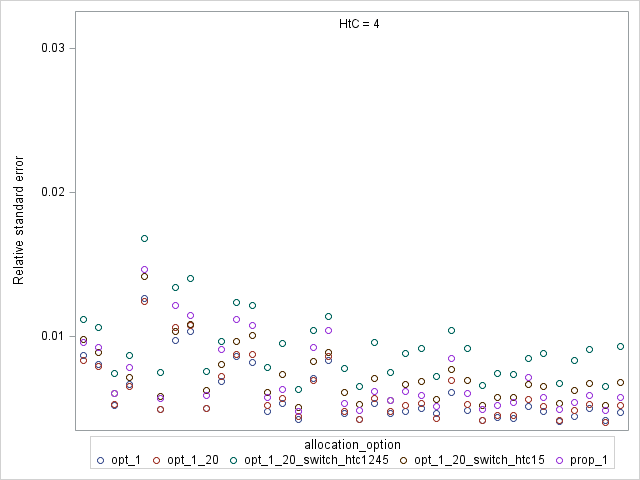
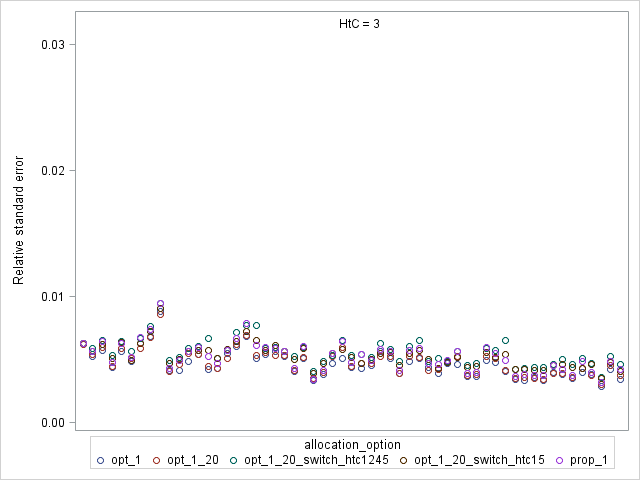
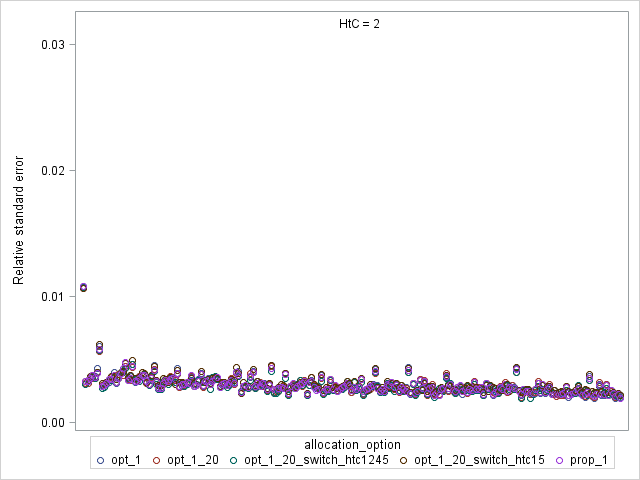
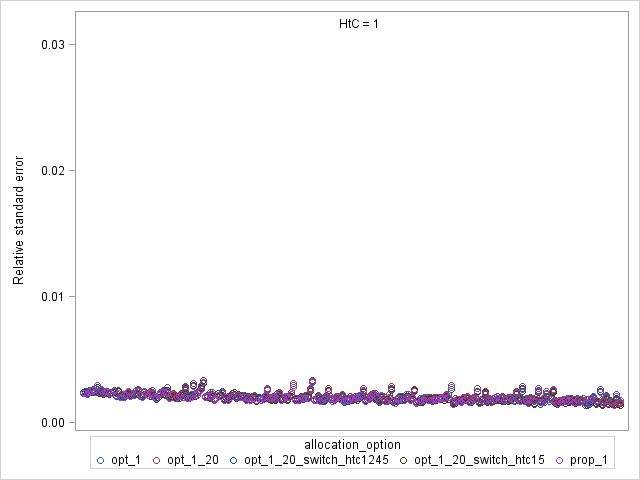
1. b)



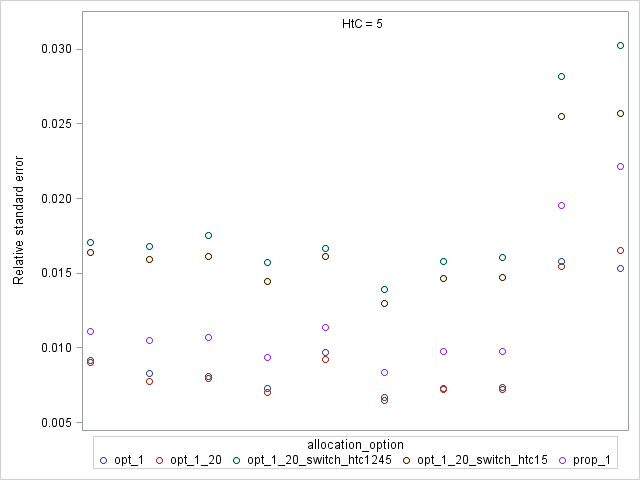
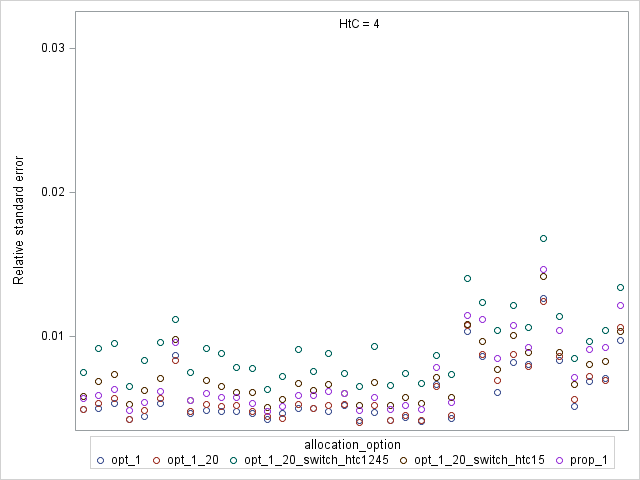
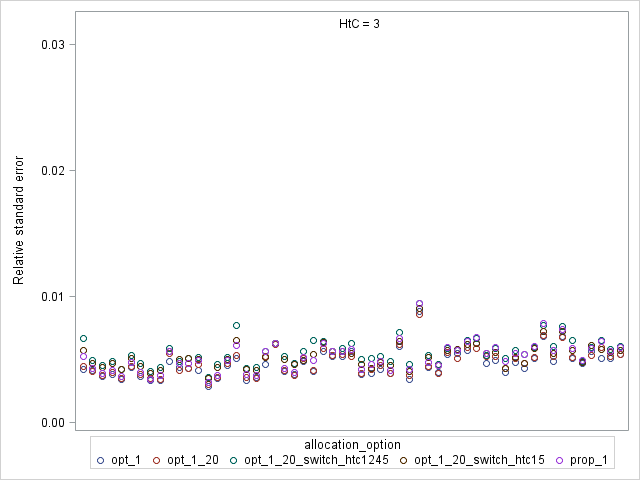
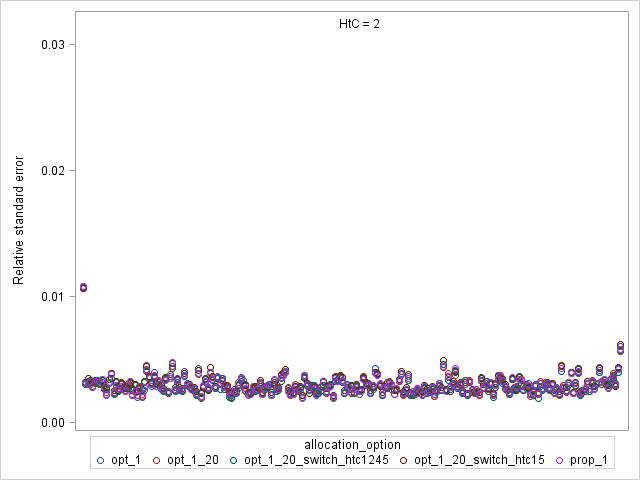
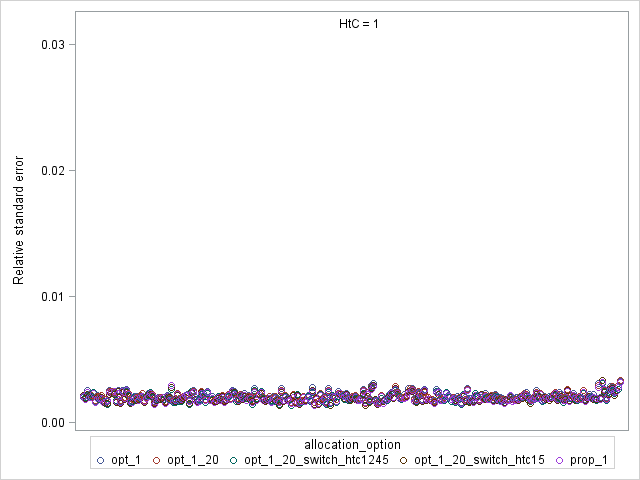
*Figure E4: RSE for LA / HtC totals for options 1 (optimal allocation) and 3 (proportional allocation) assuming CCS non-response and fitted with the logistic regression model. Scatter plot (a) ordered by RSE size in option 1.*



*Figure E5: Scatter plots of RSE for LA / HtC totals for options 1 (optimal allocation) and 3 (proportional* *allocation) assuming CCS non-response, ordered by RSE size (option 1) and displayed by HtC index category. Fitted with the logistic regression model.*



*Figure E6: Scatter plots of RSE for LA / HtC totals for options 1 to 5 assuming CCS non-response, ordered by number of PSUs within each stratum and displayed by HtC index category. All fitted with the logistic regression model.*

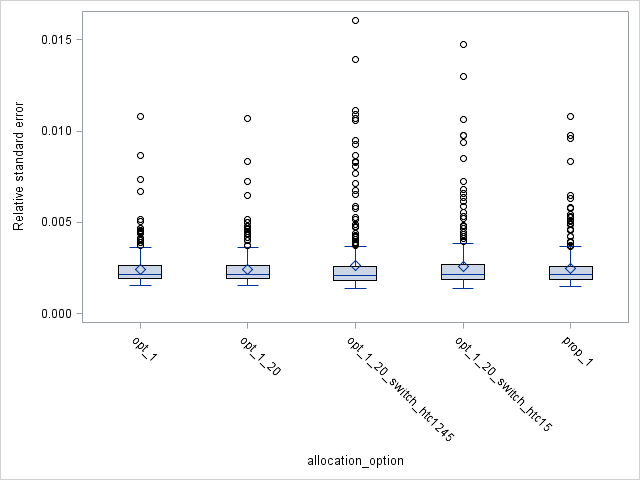
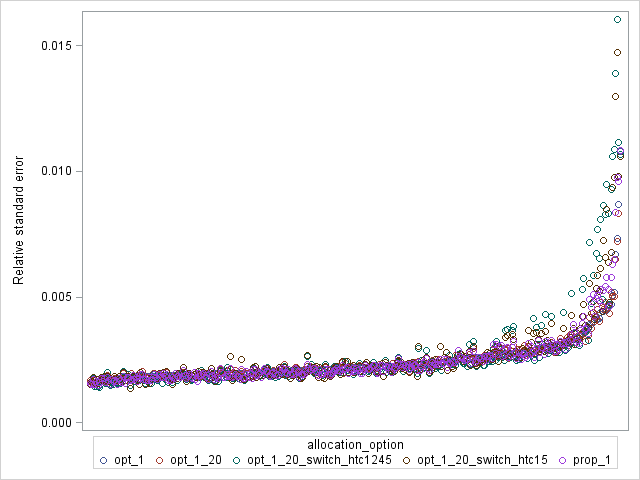


*Figure E7: Scatter plots of RSE for LA / HtC totals for options 1 to 5 assuming CCS non-response, ordered by size of standard deviation for each stratum and displayed by HtC index category. All fitted with the logistic regression model.*

*Relative standard errors for local authority totals (total of 348 points)*

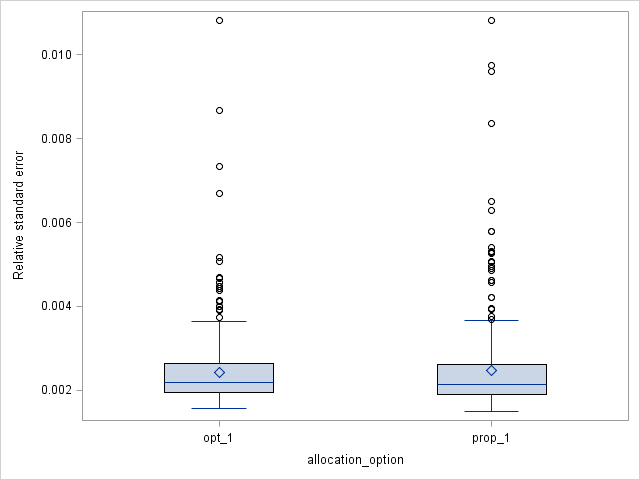
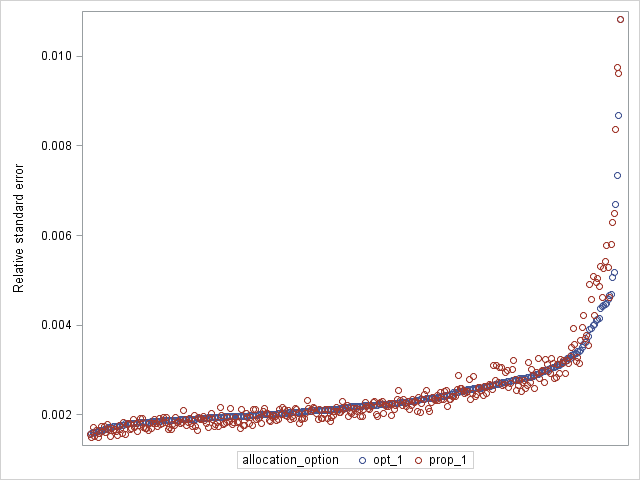
There are similar findings for LA total estimates as we see for LA / HtC estimates. Corresponding figures shown below.

1. b)



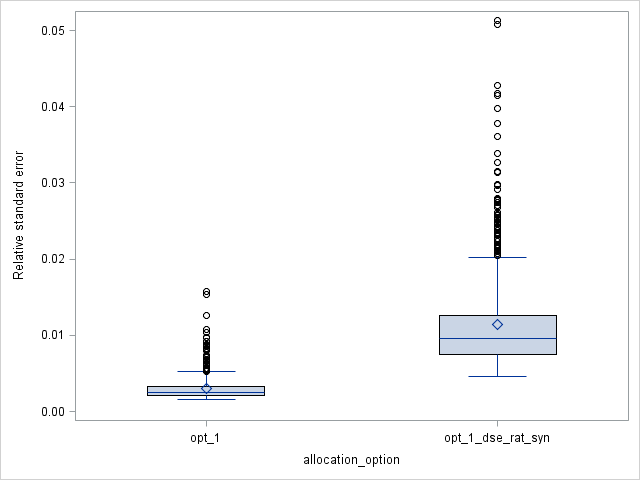
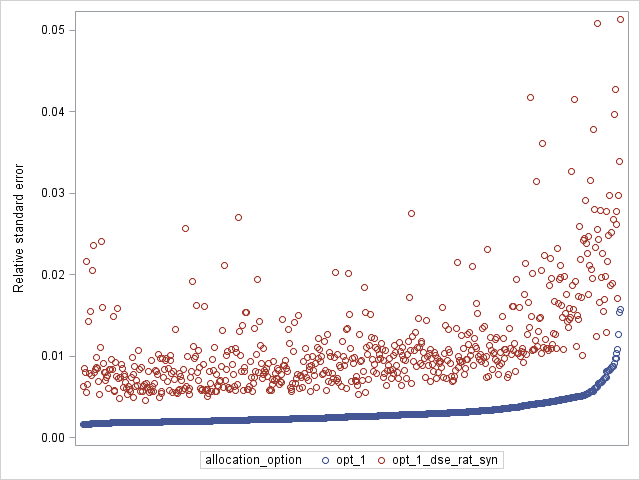
*Figure D8: RSE for LA totals for options 1 to 5 assuming CCS non-response and fitted with the logistic regression model. Scatter plot (a) ordered by RSE size in option 1.*

1. b)



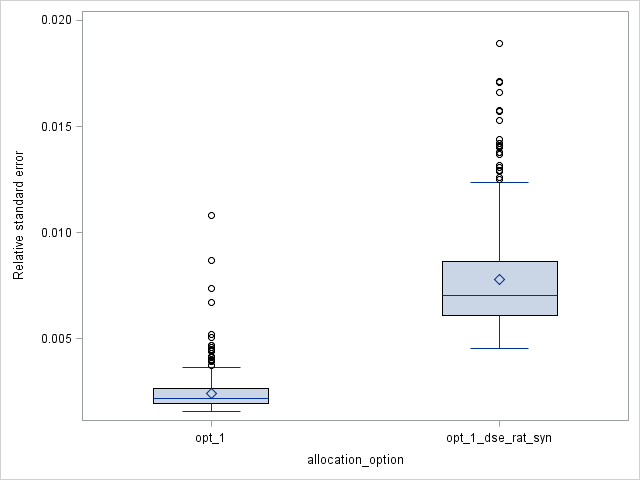
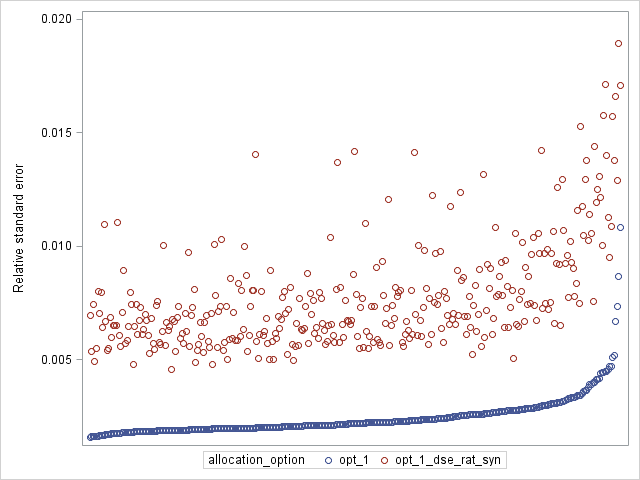
*Figure E9: RSE for LA totals for options 1 (optimal allocation) and 3 (proportional allocation) assuming CCS non-response and fitted with the logistic regression model. Scatter plot (a) ordered by RSE size in option 1.*

1. b)



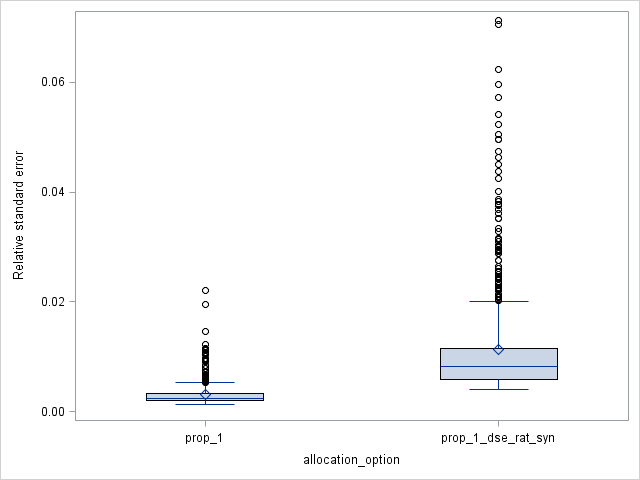
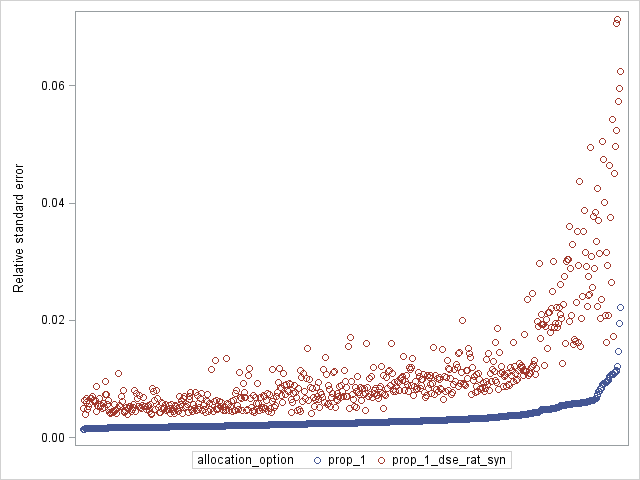
*Figure E10: RSE for LA / HtC totals for option 1 assuming CCS non-response and fitted with the logistic regression model and the dual system, ratio, synthetic estimators. Scatter plot (a) ordered by RSE size in logistic regression model.*

1. b)



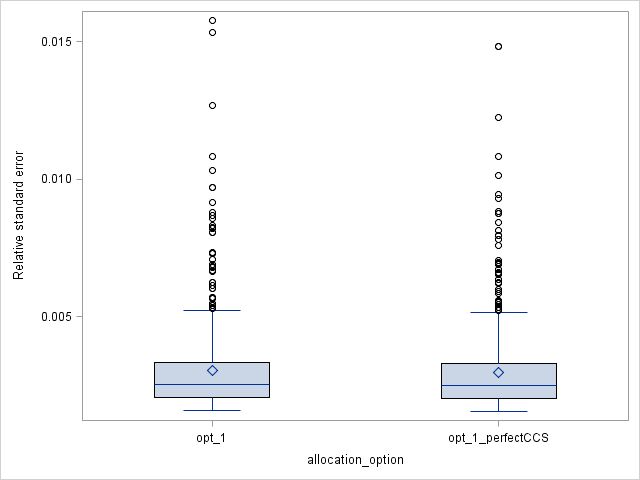
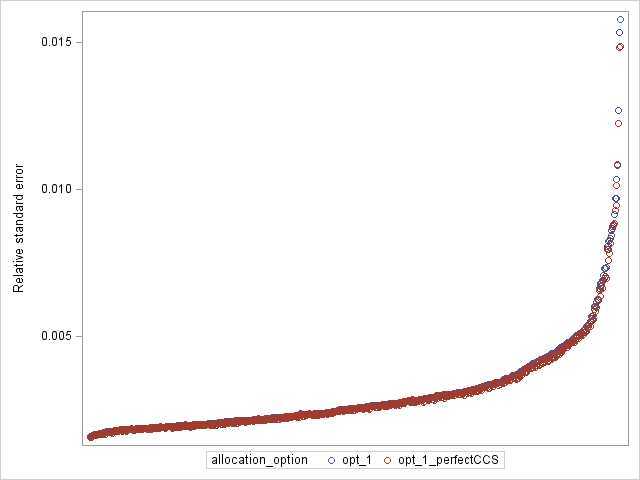
*Figure E11: RSE for LA totals for option 1 assuming CCS non-response and fitted with the logistic regression model and the dual system, ratio, synthetic estimators. Scatter plot (a) ordered by RSE size in logistic regression model.*

1. b)



*Figure E12: RSE for LA totals for option 3 assuming CCS non-response and fitted with the logistic regression model and the dual system, ratio, synthetic estimators. Scatter plot (a) ordered by RSE size in logistic regression model.*

1. b)



*Figure E13: RSE for LA / HtC totals for option 1 assuming CCS non-response and perfect CCS response and fitted with the logistic regression model. Scatter plot (a) ordered by RSE size in option 1.*