Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Scotland's Census 2022

Census Coverage Survey Sample Allocation and Reserve Sample Methodology

May 2020

National Records of Scotland

Contents

1. Introduction

1.1 High-level Summary

The Census Coverage Survey (CCS) is a voluntary survey that takes place 6 weeks after census day. It collects information from around 1.5% of people in Scotland. The CCS is used, along with Census data, to help estimate the total population of Scotland. This paper will explain how households will be chosen to take part in the CCS and grouped together to ensure there is an accurate representation of everyone in Scotland. The paper also explains how a reserve CCS sample could be used if the Census response rate was very low.

1.2 Overview of the CCS

The Census Coverage Survey (CCS) is a voluntary, interviewer led, follow-up survey that takes place 6 weeks after census day. The CCS samples approximately 1.5-2% of the population in Scotland and collects information at an individual and household level. The primary aim of the CCS is to gather age-sex data which can be used in conjunction with census data to provide population estimates. The CCS data undergoes a matching process to the collected census data and the resulting output allows us to identify the persons and households enumerated in both the census and CCS or those captured in one but not the other.

Estimation & Adjustment (E&A) applies Dual System Estimation which uses the matched CCS and Census data to estimate the number of persons or households that have been missed overall in the Census. The initial census data is then adjusted to account for these missed individuals and provide a more accurate estimation of the true population in Scotland. The CCS is therefore a crucial factor in ensuring a complete, well-rounded population count.

1.3 Overview of 2011 Methodology

In 2011, 1.5% of all households in Scotland were sampled by the CCS (~ 45,000 households and 400 Communal Establishments (CEs)) with an overall return rate of 87%.

The CCS sample design was a two stage cluster sample, stratified by Local Authority (LA) and Hard to Count (HtC) index. The HtC index is a scale of 1 (easiest to count) to 5 (hardest to count) which was created to indicate how difficult it may be to enumerate a particular geographical area based on certain demographic features. The 40% easiest to count areas are assigned as HtC 1, with the next 40% to HtC 2, 10% to HtC 3, 8% to HtC 4 and the hardest 2% to count assigned to HtC 5.

Stage 1: Selection of the Primary Sampling Unit in 2011

The first stage of sampling used Datazones[1] as the Primary Sampling Unit (PSU) with 4% of the total selected using optimal (Neyman) allocation. This sampling strategy was used to allocate the overall sample among each LA/HtC strata in proportion to the size and variance of the stratum.

In 2011, HtC levels were collapsed where there were less than 20 Datazones per HtC level. As such, when an HtC level contained less than 20 Datazones, they were moved to the next available HtC level to ensure adequate sample size for the E&A process; this was the case in all but one of the processing units in 2011.

Stage 2: Selection of Secondary Sampling Unit in 2011

Once the Datazones were selected, a set proportion of Secondary Sampling Units (SSU) were sampled from each PSU. Postcodes were utilised as the SSUs in 2011.

[1] The data zone geography covers the whole of Scotland and nests within local authority boundaries.

The second sampling stage involved the selection of 50% of the postcodes within each Datazone by a method of simple random selection.

1.4 Purpose of Document

To ensure precise and unbiased estimates, the CCS sample must provide an adequate, representative sample of the population to enable accurate estimation. This sample must be distributed across the population so as to minimise variation in the population estimates. This is achieved through statistically efficient sampling techniques while utilising a suitable sample size. There will always be a balance between increasing the sample size for improved statistical precision and the costs to conduct the survey.

This paper examines how to stratify the CCS to make best use of the sample available. Stratification by Hard to Count (HtC) and/or Local Authority (LA) are considered. It also examines the most appropriate way to allocate sample to these strata. Firstly, allocation based on 2011 response patterns is considered so as to minimise estimate variance. Then, the possibility of allocating according to 2022 Census return rates is evaluated.

This paper will also investigate a reserve sample, to be used as a contingency if there is evidence that response rates to the Census are critically low. This is based on previous research conducted by the Office of National Statistics (ONS) and their approach in the 2011 Census. While in 2021 the ONS methodological position has changed and they do not intend to boost their sample (due to the ONS new modelling approach that NRS have not adopted), the ONS are still retaining plans for a reserve sample as an emergency contingency.

The precision associated with different CCS sampling approaches and Census response rates are compared in terms of population estimate Relative Standard Error (RSE) and Confidence Intervals (CI) simulated through runs of the Estimation process with different samples.

2. Sample Design for 2022

In 2022, we aim to improve upon the design of the CCS and quality of the estimates produced in 2011, using a sample design and sample size which meets the target Key Performance Indicators (KPIs) for precision with maximum efficiency. A list of KPIs related to statistical quality are given in Appendix B. The 2011 CCS aimed to sample between 1-2% of the population. This has been used as a rule of thumb by a number of statistical agencies (ONS, Statistics Canada, Australian Bureau of Statistics) in order to maintain consistency between the balance of overall sample size and associated cost.

2.1 Core Method for 2022

In 2022 there will be a slight change in the sampling units from 2011. Instead of Datazones, the PSUs in 2022 will be Planning Areas while the SSUs will remain as postcodes. The change in the geographical aggregation from Datazones to Planning Areas is to facilitate easier enumeration and travel within the area for field force workers.

Two methods were primarily used to compare different sampling strategies. Further details on this analysis can be found in the Census Coverage Survey Sample Methodology Paper.

The Key Performance Indicator for the Estimation system is to produce 95% confidence intervals on the estimates within ±0.4% at national level. The required Relative Standard Error (RSE) to achieve this target precision is 0.204%. It was agreed by NRS Statisticians that the RSE should be less than or equal to 0.19% to increase the likelihood that the national level KPI target is achieved.

Analysis of different PSUs and SSUs was carried out under the assumption that the 2022 census response rate will be the same as in 2011 (94% response rate) and the 2022 CCS response rate will be 80%. This is 7% lower than in 2011. The reason for

this is based on the voluntary nature of the CCS and a known decline in public response to surveys.

The results of this analysis suggest that greater statistical efficiency is achieved through decreasing the level of clustering of the sample, with a minimal increase in travel time for field force workers. Selecting more planning areas (PSUs) and less postcodes (SSUs) in each planning area results in an improvement in expected precision of the estimates. It was agreed within NRS that the most statistically efficient sample would be that shown in Table 1.

Table 1: RSE value from Estimation simulations

PSU %	*SSU %*	*Estimate RSE %*	*Sample household count*
9	*17.5*	*0.1778*	*46,583*

Modelling of Field Force visits to the CCS were carried out in order to estimate the difference in travel time between different clustering strategies – the results are show in Table 2.

Table 1: Travel times as a proportion of total time worked under different sampling clustering

Sample	*Time travelling (% of total)*	*Postcodes in sample*
4% PSU 45% SSU	23.4	2479
7% PSU 25% SSU	23.9	2429
9% PSU 20% SSU	25.2	2478

Clustering did not appear to dramatically affect travel times, as the increase in travel time observed in the model is not large relative to the total travel time. This may be due to Planning Areas (the clusters) being so small that interviewers had to travel

between different clusters irrespective of how many SSUs (post codes) are selected in each PSU.

It should be noted that in the field force model there is some variation in the sample size, and correspondingly in the number of interviewers needed, so the total number of hours worked between each sample is not exactly the same.

2.1.1 Decision on core methodology

The decision taken is to have a PSU rate of 9%, a SSU rate of 17.5% and a sample size of approximately 46,583 households (there will be some variation due to the nature of sampling). This gives a RSE of 0.1778%, which is sufficient to meet national targets with a Census response rate of 94% and a CCS response rate of 80%.

2.2 Stratification and allocation strategy

2.2.1 Stratification

The purpose of the CCS is to provide good quality response data that can be matched with the Census to be used in the Dual System Estimation (DSE)[2] process in obtaining population estimates. DSE is run separately on different population strata; for this purpose the country is divided by geography (processing unit - groups of LAs) as well as by demography (by HtC). To achieve stable estimates across these strata, the CCS must sample evenly across the country.

In 2022 we aim to improve the quality of population estimates. For this reason, previous research by NRS investigated various stratification options. Preliminary analyses were conducted using an optimal allocation strategy and different sampling proportions to investigate various stratification options and the impact on statistical efficiency and associated precision of estimates, measured through RSE. The

[2] Link to DSE methodology paper
https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf

stratification options considered were LA/HtC (as in 2011), LA only, LA and altered HtC index (1-3), and no stratification.

The analysis showed that stratification by LA/HtC appeared to produce the most consistent RSE for both household and person design variables, and is therefore recommended as the most appropriate option in terms of producing precise estimates in 2022.

Stratification by these variables will also have a positive impact on the granularity of our estimates. When carrying out DSE, a sufficient sample is needed in each stratum to enable an appropriate level of precision in our estimates – if this is not achieved then strata are collapsed together. As such, by stratifying the CCS sample by both LA and HtC, we ensure that sample is evenly spread across the DSE strata. This increases the likelihood that the strata will be an appropriate size which will reduce the need for collapsing and increase the granularity of the estimates.

2.2.2 Allocation

Once appropriate strata are decided upon, a method of distributing sample amongst them must be specified. One possibility would be to split the sample evenly amongst the strata, but this would cause the smaller strata to be over represented. Another possibility would be to use proportional allocation to distribute sample according to the relative size of the strata. However, this only accounts for the size of the strata, not differences in their response characteristics. Therefore, a final option is to use optimal allocation which increases the sample both in larger areas and in those with more response variation.

The allocation strategy that was used in 2011 was Optimal (Neyman) allocation, which allocates a sample according to stratum size and variance. In 2011 a design variable was calculated, using the initial 2001 person and household data and the final post-census adjusted person and household response rates in 2001 (see Appendix A). This design variable measured variability in response patterns at a

postcode level, and reflected the magnitude of variance in relation to the average across the entire population in the 2011 census. The use of a design variable creates an outcome variable that increases proportionally to the error in Dual System Estimation (DSE), and allocates according to this and strata size.

We conducted analysis to investigate the impact of allocation method on RSE, using both optimal and proportional allocation. In this analysis the design variable was modelled on 2011 data. The RSEs where calculated using the method detailed in appendix A - 6.1 and 6.2. Results from this analysis can be found in table 3.

Table 3: Results of Allocation Analysis

		Household RSE %		Person RSE %	
Stratification	***PSU%/SSU%***	***Optimal***	***Proportional***	***Optimal***	***Proportional***
CA/HtC	**4/50**	0.164	0.166	0.080	0.080
	7/25	0.102	0.119	0.051	0.056
	9/20	0.227	0.232	0.085	0.169

The results indicate that optimal allocation resulted in a lower RSE value in comparison to proportional allocation, with the exception of a sampling proportion of 4/50% when RSE_{person} was equal for optimal and proportional (0.080). The findings suggest an increased level of precision when adopting an optimal allocation strategy, and therefore this may be a more appropriate option for use in 2022.

While this method of allocation is beneficial in allocating the sample to account for predicted response rates, the design variable is calculated using 2011 response data, which may not be reflective of response patterns in 2022. Therefore, there is an inherent risk in basing sample allocation on 2011 response patterns, particularly given that we are now adopting a Digital First approach in 2022. Mitigating factors will be discussed in subsequent sections of this paper.

2.2.3 ONS approach

ONS have explored estimation and allocation strategies, and reported on using a modelling approach using generalised linear models rather than the DSE and ratio estimation model used in 2011.[3] The logistic regression model approach favoured by the ONS is implemented at a national level to gain maximum statistical power, instead of being run separately for each strata as in DSE. Because of this change the ONS are reviewing their allocation strategy, considering proportional and optimal allocation methods, as well as a hybrid optimal-proportional allocation method. The choice of method involves some degree of trade-off between bias and variance in population totals. There is no design bias introduced in proportional allocation, whereas disproportional allocation can introduce bias unless weighted (which would cause design error) in a national estimation model such as the regression approach used.

The idea of the hybrid option 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 is a compromise strategy between proportional and optimal allocation that ensures enough sample is allocated to all hard-to-count areas, without allocating too closely to 2011 patterns of census response.

The ONS are proposing to use the hybrid optimal-proportional allocation approach as the variances are small for all allocation methods with a generalised linear model and there is little difference observed between the optimal and optimal-proportional approaches.

[3] More information on the ONS research can be found at https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/researchoutputscoverageadjustedadministrativedatapopulationestimatesforenglandandwales2011#estimation-methods
And at https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji180426

ONS analysis has found that their logistic regression approach results in smaller error rates than in the DSE model due to the increased amount of statistical power available. Because of this, optimal allocation may not provide a gain in precision that is sufficient to risk introducing additional bias into their model. These concerns are less of a consideration in Scotland given that we are continuing to use DSE. As DSE is conducted stratum by stratum, it is protected against this kind of bias that could be caused by non-proportional sample allocation or incorrect model specification. On the other hand, our RSEs are higher than would be the case if using a similar regression approach to that used by the ONS, where statistical power is shared across neighbouring geographic areas. Because of this, the decrease in error that may be gained by optimal allocation is of greater value in Scotland.

The ONS study found that that the reduction in RSE for optimal compared to proportional allocation was much greater for DSE than logistic regression estimation, which further supports our recommendation for NRS to adopt an optimal allocation strategy for our sample. However, the use of DSE based on design variables reflecting 2011 response patterns runs the risk over weighting a specific area (strata) of the country in estimation. This in itself is not a significant issue, as each stratum has DSE applied individually. However, if subsequent collapsing occurs, this could result in over-allocation of large proportions of our sample, which could impact the quality of our estimates.

When there were similar expected and realised results, optimal allocation was found to produce slightly increased precision when compared to proportional allocation, for both DSE and logistic regression estimation. However, conversely, when allocation is sub-optimal (if predicted results vary drastically from actual results), RSE is higher for optimal allocation than proportional allocation. Therefore, by optimally allocating our sample based on design variables that model variation in 2011 response patterns, if response patterns in 2022 are not consistent with those of 2011, then there will likely be an increase in RSE and an impact on the quality of our estimates.

2.2.4 Decision on 2022 Methodology

Based on findings of research described earlier in this paper as well as evaluation of the ONS' approach, stratification is proposed by LA and HtC index with sample allocated to these strata using optimal allocation for the majority of the CCS sample in Scotland. However consideration must be given to the risk of over calibrating allocation to the 2011 response patterns, which could reduce the quality of the overall population estimates if response patterns change.

3. Flexible sample

Given the fact that Scotland will be utilising a digital first collection approach for the Census in 2022 for the first time, there is a concern that optimal allocation to 2011 response rates might be inappropriate. It might be that due to the push to online capture, digitally excluded areas have lower response rates than predicted based on 2011 data. It is also possible that other changes in population demographics might alter response patterns. One way of mitigating the risk of large variations from expected response is to reserve some of the sample for allocation once response rates to 2022 are known.

The proposal is to hold back some of the original sample size as a flexible sample that can be allocated according to trends in 2022 return rates. This sample would be allocated no later than five weeks after Census day to allow sufficient time for the sampling areas to be divided up amongst interviewers and operational planning and organisation to take place. In order to conduct the necessary analysis for sample selection and allocation by this time, analysis of return rates must start four weeks after Census day.

The Census data collection will not be complete in time to conduct the necessary analysis for CCS sample selection, so the final return and response rates for the Census will not be known. Allocation will instead be made based on projected final return rates for whole strata (as opposed to the more granular household level response rate information from 2011 that drives the main CCS sample allocation).

This paper assumes that it will be possible to predict the general pattern in final Census return rates with reasonable accuracy from the return rates four weeks after Census day. A method for doing this is not presented in this paper.

The following section analyses the impact of this approach in two scenarios: one in which Census response rates are largely driven by digital exclusion index (DEI)[4]; and another with response rates randomised for each estimation strata. Both scenarios had an overall response rate of approximately 91%. The precision of estimates (expressed as confidence intervals and RSEs) associated with the following three options is measured in each scenario:

- 100% optimal allocation according to the 2011 design variable (DV)
- 80% optimal allocation to the 2011 DV and 20% 'flexible' - using the strata level response rate variation
- 100% 'flexible' allocation – using the strata level response rate variation

3.1 Flexible Sample Simulation Modelling

3.1.1 Methodology

Optimal allocation in the majority of the sample is conducted based on a design variable that captures postcode level variations in the response rate around the average. Therefore, the design variable will tend to be larger for, and thus more sample allocated to, strata that have either lower response rates or a higher level of response rate heterogeneity over their area. At the time the flexible sample is to be allocated there will be insufficient information at this granularity to create such a design variable for 2022 responses. The best information that is likely to be available will be the return rates by strata four weeks after Census day, that can be used to predict final return rates by strata.

[4] The Digital Exclusion Index (DEI) is a ranked list of planning areas which is ordered by how many people in the area in terms of proportion are predicted to be digitally excluded (lacking internet access or digital skills). Planning areas are categorised on a scale of 1 to 5, with 1 being the least digitally excluded and 5 being the most digitally excluded.

The proposed methodology for the flexible sample therefore is to allocate a portion of the sample according to the return rates at the stratum level. To do this the standard deviation in the binomial return rate was calculated for each stratum.

1. $\theta_h = \sqrt{\frac{p_h(1-p_h)}{N_h}}$

   Where θ is the standard deviation, p is the proportion of houses returning, (1-p) is the non-return rate and N is the number of houses expected to respond in each stratum h.

The standard deviation was then used alongside stratum size to calculate the optimal allocation n_h for each stratum

2. $n_h = n * \frac{(N_h * \theta_h)}{\sum (N_h * \theta_h)}$

   Where n is the total sample.

This approach will capture the stratum by stratum variations in response rate, but will not capture the level of heterogeneity in response rate within individual strata (for a given stratum level response rate, the underlying response patterns could be homogenous or heterogeneous across the stratum). Therefore it may not be ideal to allocate 100% of the sample according to the 2022 response rates. For this reason three different options where investigated:

- 100% optimal allocation according to 2011 response rates
- 80% allocation to 2011 response rates and 20% to 2022 response rates
- 100% allocation to the 2022 response rates

The census response rates modelled where based on the 2011 Census returns, but modified for the two scenarios (high impact of digital exclusion and randomised response by estimation strata).

DEI modified Census response rate

The starting point was the current predicted response rate of 91.2% (86.2% + 5% due to Communications marketing) from NRS modelling work. The predicted response rates by HtC can be seen in table 4.

Table 4: Starting response rates by HtC

HtC	*Response*
1	95.1
2	90.7
3	85.3
4	83.9
5	79.6

A modification for each level of the DEI was applied, this can be seen in table 5.

Table 5: Size of modifier applied to each DEI level

DEI	*modifier*
1	4
2	3
3	1.6
4	-3
5	-10

This then results in the following response rates by stratum:

Table 6: Final scenario response rates by HtC and DEI

	DEI				
HtC	1	2	3	4	5
1	99.1	98.1	96.7	92.1	85.1
2	94.7	93.7	92.3	87.7	80.7
3	89.3	88.3	86.9	82.3	75.3
4	87.9	86.9	85.5	80.9	73.9
5	83.6	82.6	81.2	76.6	69.6

The response rate for the CCS was set at 80% for all demographics and strata.

Random modified Census response rate

The census estimates are stratified by a combination of Estimation Areas (EA) and HtC groups – these strata area called estimation area sub-groups. Each estimation area is made up of a combination of LAs which are likely to have similar response rates. During Census follow-up, equalising response rates within estimation area subgroups is prioritised.

In this scenario response rates for each estimation area strata where assumed to vary strongly from the 2011 response pattern, but the strata remain homogenous. For each estimation area the response rate was randomised, with the constraint that the final overall response rate should equal approximately 91%.

The effect of the flexible sample on population estimate precision was modelled via simulation for the three options outlined above (section 3), in the two scenarios – DEI response rate and randomised response rates for EA strata - alongside a baseline using 2011 response patterns and 100% allocation based on the 2011 design variable. Simulations involved executing 500 runs of DSE under differing synthetic CCS and Census response data sets (resampling from the 2011 Census population so as to achieve the target response rates) then examining the variability in the

resulting population estimates to obtain confidence intervals. Further details of this method can be found in appendix A – 6.3.

3.1.2 Results

The results of the simulations can be found in figure 1. For the baseline with CCS allocation according to the design variable based on 2011 response patterns and Census responses following projected response patterns, the confidence interval was 0.362% - meeting the KPI for precision (0.4%). The new 'flexible' sample allocation method also performed well in this scenario, with CI of 0.358% and 0.360% for the 20% flexible and 100% flexible sample allocations respectively.

The first scenario under investigation – a strong influence of DEI on Census response rates – caused some problems in meeting the KPI for precision. When the CCS sample was 100% allocated according to 2011 response patters (using the design variable) the simulation showed a CI of 0.452%, above the 0.4% KPI target. The 20% flexible allocation based on DEI response rates had a similar precision with a CI of 0.455%. The 100% flexible allocation improved the CI to 0.421%, although the is still above the KPI.

The second scenario – randomised response rate within EA strata – was also problematic for the standard 2011 DV allocation (CI = 0.417%). This improved to just within the KPI for the 20% 'flexible' (CI = 0.398%). The 100% flexible sample performed the best for this scenario with a precision CI of 0.365%.

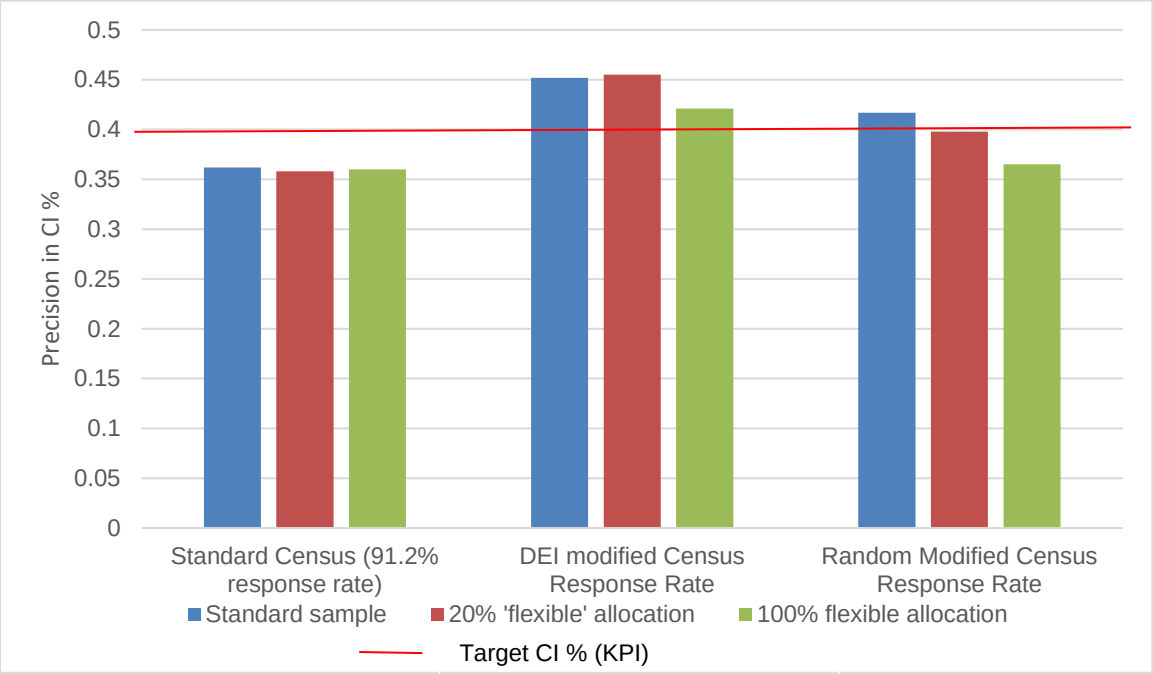

Figure 1: Modelled Confidence Intervals (CI) for the various scenarios

3.2 Discussion

The results clearly demonstrate the issue of over allocation to the 2011 response patterns. In a scenario in which Census response rates are largely driven by digital exclusion, precision dropped and failed to meet the KPI for the allocation based on the 2011 design variable. Similarly, when response rates were randomised the precision associated with the standard allocation were above the KPI. This is of course only two ways in which response rates might deviate from 2011 patterns, but other response deviations (if of the same magnitude) should cause similar outcomes.

The results indicate that there is a benefit to using return information from 2022 in allocation. When the whole sample was allocated using the 'flexible' method, according to the 2022 Census response rates, there was an improvement in precision in all scenarios. Again, it seems likely that using a flexible sample would be beneficial in any situation where there was deviation from the 2011 response patterns.

One caveat of this method is the simulation method does not adequately recreate the within strata heterogeneity of response rates. The 2011 design variable for the standard allocation tracks variations at the post code level and therefore excels at picking up this kind of variation. On the other hand, the best information that will be available when the sample is allocated will be 2022 Census response rate by stratum. Therefore the flexible sample allocation based on this information assumes the response rates are homogenous within strata. Because of this it is arguable that the simulation method favours the flexible method, although it should be noted that the 2011 design variable approach will only perform well where within strata response rates mirror the 2011 patterns.

It seems then that there is a trade-off to be made between the two methods – over allocation to the 2011 design variable discards information about 2022 response rates, while over allocation to 2022 response rates discards information about response rate heterogeneity within strata that will likely still be relevant. Therefore while the results suggest that 100% flexible allocation should be preferred in all situations, this is likely to be an over correction in reality.

The decision on how much sample should be allocated to each will depend on the extent of the variation from 2011 that is observed in 2022. If there is no clearly significant deviation or it is marginal, it may be that optimum results are achieved with allocation to 2011 alone. On the other hand if there is very little similarity to 2011 in the response pattern, allocation to 2022 alone may be suitable. Further work is needed to devise exactly what set of circumstances should trigger allocation to the 2022 observed response rates; how much of the CCS sample should be allocated; and to fine tune methodology to better simulate within strata variance.

A further question is how best to predict the final stratum by stratum response rates for 2022 based on return rates 4 weeks after Census day. One option would be to simply allocate based on the return rates at that point. This would assume that, while the return rates would increase over the final 2 weeks, the stratum by stratum

relative differences would not. However, field force modelling suggest that the rate of change in the return rates will not be the same for all strata. A better option therefore might be to project the final response rates from trends in each stratum, or to use expected trajectory from the NRS field force model. Further work will be required to develop the best method.

4. Reserve sample

As mentioned in section 2.1, our sample size modelling and subsequent decision is based on a Census response rate of 94% - the KPI for response rates (see appendix B) and the rate in 2011. However, it is reasonable to consider that response rates will be lower than this, and to build in a contingency plan based on that. NRS modelling work suggests we should be expecting a response rate of 86.2% (not including the effect of Communications marketing). It is also possible that there is some unforeseen circumstance that causes either a very low response rate across the nation, or a drop in a specific area. Candidates for this could be critical technical failure or a natural disaster.

The rest of this paper will explore the option of putting in place plans for a 20% reserve sample that would be activated in a contingency situation if Census return rates are very low. The impact of this extra sample compared to the standard sample will be examined in the following scenarios via simulation modelling runs:

A. A national response rate of 86.2% stratified by HtC
B. A national response rate of 70% stratified by HtC
C. A national response rate of 90% stratified by HtC, with a 40% response rate in one Local Authority (Glasgow) and the additional 20% allocated only there

4.1 Reserve Sample Simulation Modelling

4.1.1 Methodology

The effect of the 20% reserve sample on population estimate precision is modelled via simulation under the three scenarios outlined above. This is achieved by executing 500 runs of DSE under differing synthetic CCS and Census response data sets then examining the variability in population estimates to obtain confidence intervals. Further details of this method can be found in appendix A – 6.3.

The three scenarios use differing assumptions about Census response and CCS sample allocation as inputs. In all scenarios the CCS response rate is set at 80%; our baseline planning assumption.

Scenario A: A national response rate of 86.2% stratified by HtC

Based on recent NRS modelling, the following response rates' by hard to count index (HtC) were used. The overall Census response rate was 86.2%.

Table 7: Response rates by HtC for Scenario A

HtC	*Response rate*
1	90.1%
2	85.7%
3	80.3%
4	78.9%
5	74.6%

Scenario B: A national response rate of 70% stratified by HtC

Based on a reasonable assumption around a critically low response rate to the Census, a 70% response rate was modelled, stratified by HtC in the same relative proportions as under Scenario A.

Table 8: Response rates by HtC for Scenario B

HtC	*Response rate*
1	74.0%
2	69.6%
3	64.2%
4	62.8%
5	58.5%

Scenario C: A national response rate of 90% stratified by HtC, with a 40% response rate in Glasgow

Based on a scenario which assumes a good national response rate to the Census but a catastrophically low response rate in one council area (Glasgow in this example) the following response rates were used across the board and in Glasgow.

Table 9: Response rates by HtC and LA for Scenario C

HtC	*Rest of Scotland*	*Glasgow*
1	93.9%	43.9%
2	89.5%	39.5%
3	84.1%	34.1%
4	82.7%	32.7%
5	78.4%	28.4%

All of the additional 20% sample was allocated to Glasgow in this scenario. That constitutes an additional 162 planning areas, or 15% of all Glasgow planning areas, leading to a total sample of 24% of all Glasgow planning areas (when the CCS baseline sample of 9% is added in).

4.1.2 Results

Scenario A and B

The modelling results under scenario A and B are displayed in figure 2 above. As this shows, under every scenario run the Confidence Interval (CI) did not meet our precision KPI of 0.4% at a national level. This is due to the decrease in current assumptions for response rate compared to our KPI for response rates. When the reserve sample was added in, the CI improved in both scenarios. Interestingly, the reserve sample had the largest relative impact in the critically low response rate scenario compared to the current field force model forecasts.

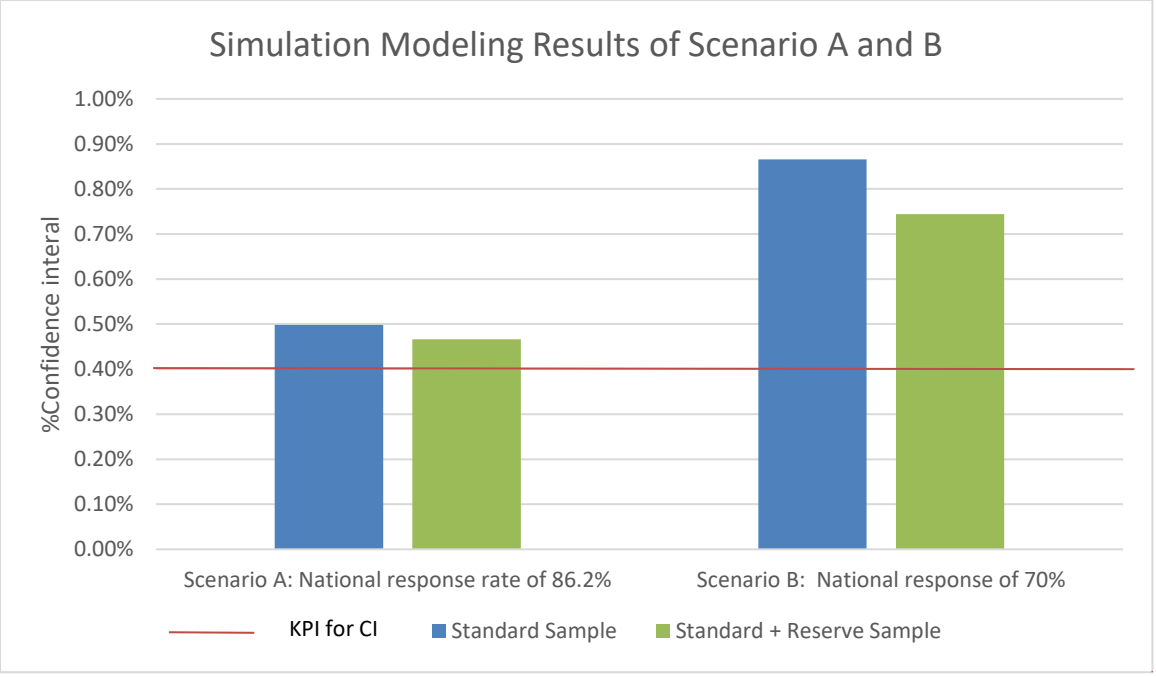

Figure 1: Chart of Confidence Intervals under scenarios A and B given both a standard sample and standard + reserve sample (red line indicates our target KPI for precession)

Scenario C

Table 10 shows the modelling simulation results under scenario C in which we assume a fair national response rate (90%) but a catastrophically low response rate in Glasgow (40%). Under this scenario we see a failure to meet the national KPI for both the standard and targeted reserve sample, but a marked improvement with the

targeted reserve. For Glasgow itself there is a failure to meet the local authority target (3%) under the standard sample, but an improvement to within the target for the standard + targeted reserve sample.

Table 10: Table of modelled Confidence Intervals for the Nation and Glasgow (Our KPIs are 0.4% for National estimates, and 3% for local authority).

CCS sample	National CI	Glasgow CI
Standard	0.542%	3.77%
Targeted reserve	0.480%	2.76%

4.2 Discussion

The activation of the reserve sample shows marked improvements in the precision KPIs in all scenarios. Of particular note, the reserve sample makes a larger relative improvement in the more seriously poor response rate scenarios B and C.

The best results in terms of meeting KPIs are when the CCS reserve is used to target a specific area with catastrophic response rate. This is due to the large concentration of sample that can then be used in that area.

5. Conclusion

This paper explores the method of stratification and allocation to use for the 2022 CCS. First, Section 2 reviewed previous research at the NRS and elsewhere around the core CCS stratification and allocation strategy. Then section 3 looked at new research on a method, the flexible sample, for mitigating some of the risks in the proposed allocation strategy. Finally, section 4 looked at the use of a reserve sample which could be activated if Census return rates were very low.

5.1 Core Stratification and Allocation strategy

As DSE is conducted on individual PU/HtC strata, stratification of the CCS is required to ensure that there is sufficient sample in each of these strata to conduct the analysis. Furthermore, using both geographical and socio-economic strata helps ensure that areas of potentially differing response rates all have sufficient weight. For this reason the proposal is to keep the 2011 strata for the CCS.

Previous work at NRS has looked at how best to allocate the CCS sample to strata. This work, along with research from the ONS, indicated that optimal allocation based on a design variable that tracks variation in response rate patterns in 2011 is the most efficient allocation method.

Therefore, we propose stratification by LA and HtC index with sample allocated to these strata using optimal allocation for the majority of our CCS sample. However consideration must be given to the risk of over-allocating strata based on 2011 response patterns, which could introduce bias and reduce the quality of our overall population estimates.

5.2 Flexible sample

The utilisation of a flexible sample is suggested to mitigate the risks in allocation to 2011 response rate. The proposed method is to leave allocation of a portion of the CCS sample until 4 weeks after Census day, when preliminary Census response rates for 2022 will be available. At this point non-response follow up will not be complete, so the final return and response rates will not be known, so allocation will instead be made based on projected final return rates for whole strata (as opposed to the more granular household level response rate information from 2011 that drives the main sample allocation).

Analysis was conducted on the ability of this method to compensate for two scenarios: one in which Census response rates are largely driven by digital exclusion index (DEI);

and another with response rates randomised for each estimation strata. Both scenarios had an overall response rate of approximately 91%. The precision of estimates (expressed as confidence intervals and RSEs) associated with the following three options is measured in each scenario:

- 100% optimal allocation according to the 2011 design variable (DV)
- 80% optimal allocation to the 2011 DV and 20% 'flexible' - using the strata level response rate variation
- 100% 'flexible' allocation – using the strata level response rate variation

In these two scenarios, utilising 100% allocation to 2011 caused the modelled estimate CI to rise so the precision KPI was not met. When 20% of the sample was allocated to with 'flexible' methodology, the CI also exceeded the KPI in the DEI scenario, but improved to within the KPI target in the random response scenario. However, when 100% of the sample was allocated with the flexible methodology, the CI met the KPI target in both scenarios.

While this suggests that the new methodology should always be preferred, data availability and methodological concerns suggest there is a trade-off to be made between the two methods – over allocation to the 2011 design variable discards information about 2022 response rates, while over allocation to 2022 response rates discards information about response rate heterogeneity within strata.

The proposal therefore is to hold 20% of the sample for allocation once preliminary return rate information is available 4 weeks after Census day. At this point a decision can be made on whether to allocate sample according to 2011 response rates or 2022 response rates. In an extreme scenario with a very large deviation from 2011 response rates a decision might also be taken to re-allocate some of the core sample to 2022 return rates.

Further work is needed to define the process by which this decision would be taken, and how to predict final return rates based on the return rates 4 weeks after Census day.

5.3 Reserve Sample

This paper explores the option of putting in place plans for a 20% reserve sample that would be activated in a contingency situation if Census return rates are very low. Three scenarios where modelled to examine the impact of this reserve sample in various situations:

A. A national response rate of 86.2% - based on recent NRS modelling work
B. A national response rate of 70% - a critically low response rate that might be expected if there was a large scale technical error or some other factor affecting response rates nationally.
C. A national response rate of 90%, with a 40% response rate in Glasgow and the additional 20% reserve sample allocated only there – a catastrophically low response rate in a single LA, with otherwise good response rate elsewhere, is what we might see if there was a natural disaster or some other regional crisis.

The activation of the reserve sample shows marked improvements in the precision KPIs in all scenarios. Of particular note, the reserve sample makes a larger relative improvement in the more seriously poor response rate scenarios B and C.

The best results in terms of meeting KPIs are when the CCS reserve is used to target a specific area with catastrophic response rate (Glasgow). This is due to the large concentration of sample that can then be used in that area.

It is therefore proposed to have a 20% reserve sample that can be activated in the event that Census response rates are critically low, either at the national level or regional level.

6. Appendix A

6.1 Calculation of Design Variable

To look at the effectiveness of the sample design, there are two main design variables used in this analysis: one for household response rates (Z_h) and one for the individual level (Z_i) (1). Where the design variable *Z* is the difference between the 2011 post-adjusted census counts (*Y*) and the product of the initial, unadjusted, 2011 counts (X) and the ratio (*R*); R is a ratio of the summed pre-adjusted and post-adjusted counts across all postcodes (2). The script *i* is the total count of individuals and *h* is the household count across all postcodes respectively (Brown, 2011).

$$Z_p^i = Y_p^i - R^i X_p^i$$

(1)

$$Z_p^h = Y_p^h - R^h X_p^h$$

$$R^i = \sum_{p=1}^{P} Y_p^i / \sum_{p=1}^{P} X_p^i$$

(2)

$$R^h = \sum_{p=1}^{P} Y_p^h / \sum_{p=1}^{P} X_p^h$$

These two design variables reflect the modelled variability in the 2011 census coverage at the postcode level. Our aim for 2022 is to minimise this variance. The initial 2011 person and household data and the final post-census adjusted person and household response rates for 2011 were used in the creation of these design variables. Within the context of this study, these design variables allow for the investigation of changes to the PSUs and SSUs of the clustering models by acting as a proxy for the variation of our estimates (Brown, 2011). The design variables were also utilised to conduct the optimal allocation of the sample in the first stage of sampling. The use of design variables was a key component in ascertaining the level

of improvement in the statistical design achieved through varying the sampling proportions (Brown, 2011).

6.2 Analysis of Relative Standard Error

The improvement in the estimates was determined by evaluating their expected relative standard error (RSE) for the different cluster proportions in comparison to the values obtained using the 2011 cluster values. The RSE value (3) provided a measure of the variability of population estimates and was derived via statistical analyses using the household and individual design variables. The RSE in this case was based on the modelled variability of the population divided by the total population estimate (T), a lower resulting RSE indicated an improvement in the design.

$$\%RSE = \frac{\sqrt{\{V\,(\hat{T}-T)\}}}{T} \times 100 \qquad (3)$$

The variability of the population estimate based on the 2-stage clustered sample was determined using the equation of variation outlined in Brown et al. (2011) (4):

$$V\left(\hat{T}^i - T^i\right) = \sum d \left\{ \frac{N_d^2}{n_d} \left(1 - \frac{n_d}{N_d}\right) \sigma_d^2 + \frac{N_d}{n_d} \sum o \in d \left(1 - \frac{m_{do}}{M_{do}}\right) M_{do}^2 \frac{\sigma_{do}^2}{m_{do}} \right\} \qquad (4)$$

Where M_{do} is the total number of postcodes in each HtC and planning area (o) and m_{do} is the number of postcodes sampled. Additionally N_d and n_d are the number of postcodes in the overall population and of the sample in each HtC respectively (Brown, 2011). The equation models the variation within the clusters of the sample estimates and the variation of the design variable across the population. It accomplishes this by respectively calculating the variance of the design variable for each postcode within the sample clusters (σ_{do}^2) (5) and the variability across the totals of the clusters (σ_d^2) (6). Where Z_{dop}^i is the total of the postcode design

variable within each Planning Area and Z_{do}^{-i} is the mean of the design variables for the postcodes within each Planning Area (5). Further to this, Z_{do}^{i} is the total of the postcode design variables for each HTC and Planning Area and Z_{d}^{-i} is the means for the design variables for the postcodes within each HtC and Planning Area (6). The clusters are examined at HTC stratification level (*d*) (Brown 2011). This results in the estimated variances of the population, which are then used in the equation to determine the overall variance between the two, taking into account the population estimates.

$$\sigma_{do}^{2} = \frac{1}{M_{do}-1} \sum p \in od \ (Z_{dop}^{i} - Z_{do}^{-i})^{2} \quad (5)$$

$$\sigma_{d}^{2} = \frac{1}{N_{d}-1} \sum o \in d \ (Z_{do}^{i} - Z_{d}^{-i})^{2} \quad (6)$$

These variance values were calculated for each of the 8 tentative cluster proportions, the square root of these values were then used to calculate the standard error relative to the population estimates for the individual and household design variables respectively (1). In order for the design to show improvement the overall variability of the sample design variable in relation to the population should be smaller than it was in 2011.

6.3 Methodology for Estimation Simulations

A simplified simulation of 2011 methodology was used in order to examine the effects of the CCS sample selection method on the estimates produced. The adjusted 2011 census was used to create a synthetic CCS sampling frame. The response rate within the synthetic CCS could be varied by selecting the desired proportion of households to be included in the frame.

The numbers of people in the pre-adjusted 2011 census and the synthetic CCS were aggregated within each postcode by age-sex group. This dataset can then be queried to pull out the postcodes selected through the different sample methods to

be tested, which then are aggregated by age-sex group and hard to count index within each Processing Unit (using the same groupings as used in 2011) to calculate estimates through DSE, and produce a scaling ratio between the estimates and original census count within sample areas.

These ratios were then applied to the overall population, again stratified within each Processing Unit by age-sex group and hard to count index, giving the estimates for the overall population. Estimates were only calculated by age-sex group, and no additional correction methodologies were used in the simulation. To calculate Local Authority estimates, a synthetic estimator approach was used, applying the DSE ratios for the Processing Unit to the original census count of each Local Authority separately.

From the 500 different estimates produced for each of the 500 replicate CCS samples, the variance of the average and corresponding RSE can be calculated. In cases with lower than 100% response rate, 500 different replicates of selecting which households were responding were used.

7. Appendix B

Measures of success for Scotland's Census 2022 objectives, as at November 2019[5].

How we will achieve high quality results?	How will we measure success? (Level 1 Key Performance Indicators (KPIs)[1] and acceptance levels)
We will maximise our overall person response rate	Person response rate[2] of at least 94%
We will ensure a minimum level of response with every local authority in Scotland	Person response rate in every council area of at least 85%.
We will maximise the accuracy of our national population estimates	Variability[3]: national estimates will achieve 95% Confidence Intervals (CI) +/- 0.4%; Bias: < 0.5%
We will maximise the accuracy of our local authority population estimates	Variability[4]: council area estimates will achieve 95% CI +/- 3%
We will minimise the non-response to all mandatory questions	Achieve or exceed target non-response rates for all mandatory questions
Our data will demonstrate high agreement rates with post coverage quality surveys	Agreement rates of at least XX%[5] achieved for all questions
All national and local authority level results for each main release will be assessed by a quality assurance panel	Undertaken with no residual issues remaining
We will publish details of methods and full details of all our data quality indicators	Published on our website
We will publish the results of an independent methodology review	Positive review published.
We will maintain our National Statistics Accreditation	Accreditation maintained throughout

1. Lower-level KPIs may sit below individual Level 1 KPIs.
2. Precise measure for person response rate to be defined.
3. This target is under review.
4. This target is under review.
5. Precise measure for agreement rate to be defined.

[5] As found in Scotland's Census 2021 Statistical Quality Assurance Strategy
https://www.scotlandscensus.gov.uk/documents/Statistical%20Quality%20Assurance%20Strategy.pdf

8. Appendix C

List of Acronyms

CCS	Census Coverage Survey
CE	Communal Establishment
CI	Confidence Interval
DEI	Digital Exclusion Index
DSE	Dual System Estimation
DV	Design variables
EA	Estimation Area
E&A	Estimation and Adjustment
HtC	Hard to Count Index
KPI	Key Performance Indicator
LA	Local Authority
NRS	National Records of Scotland
ONS	Office of National Statistics
PSU	Primary Sampling Unit
RSE	Relative Standard Error
SSU	Secondary Sampling Unit

Geography Definitions

Data Zone	The data zone geography covers the whole of Scotland and nests within local authority boundaries.
Digital Exclusion Index	The Digital Exclusion Index is a ranked list of planning areas which is ordered by how many people in the area in terms of proportion are predicted to be digitally excluded (lacking internet access or digital skills). Planning areas are categorised on a scale of 1 to 5, with 1 being the least digitally excluded and 5 being the most digitally excluded.

Hard to Count Index	The Hard to Count index is a scale of 1 (easiest to count) to 5 (hardest to count) which was created to indicate how difficult it may be to enumerate a particular geographical area based on certain demographic features.
Local Authority	Local Authorities are the 32 council areas within Scotland.
Planning Areas	Planning Areas are geographic areas built from groups of postcodes and averaging between 200-400 residential addresses. They nest within Local Authorities.

9. References

Brown, J., Abbott, O., & Smith, P.A. (2011). Design of the 2001 and 2011 Census Coverage Surveys for England and Wales, *Journal of the Royal Statistical Society.* 174(*4*), pp. 881-906.

Castaldo, A (2018a). 2021 Census Coverage Survey Design Strategy. Version 4. *The Office of National Statistics.*

Castaldo, A. & Nikolakis D. (2018b). Assessing the use of an Address Based Design for the 2021 Census Coverage Survey. *The Office of National Statistics.*