**Coverage Estimation Modelling – Methodology Update and Implementation Report**

Bar

**Contents**

[1. Introduction 4](#_Toc164867471)

[*1.1* *Background* 4](#_Toc164867472)

[*1.2* *Lower than expected response rates* 4](#_Toc164867473)

[2. Change of DSE model 5](#_Toc164867474)

[2.1 Under coverage in 2011 5](#_Toc164867475)

[2.1.1 Stage 1 – Dual-system estimation 5](#_Toc164867476)

[2.1.2 Stage 2 – Estimation area population estimates 6](#_Toc164867477)

[2.1.3 Stage 3 – Synthetic LA estimation 6](#_Toc164867478)

[2.2 Logistic regression approach 7](#_Toc164867479)

[2.3 Advantages of change in model approach 8](#_Toc164867480)

[2.3.1 Reduction in heterogeneity bias 8](#_Toc164867481)

[2.3.2 Reduction in LA level bias due to synthetic estimator 8](#_Toc164867482)

[2.3.3 Lower estimate random error 9](#_Toc164867483)

[2.3.4 Possibility of using operational data in model 9](#_Toc164867484)

[2.3.5 Methodologically streamlined – preliminary low level estimates early 9](#_Toc164867485)

[2.4 Challenges to implementation 9](#_Toc164867486)

[3. Use of administrative data to enhance population estimation (under coverage model) 9](#_Toc164867487)

[3.1 Population precision and the impact of admin data enhancement on error 11](#_Toc164867488)

[3.1.1 Derivation of relative route mean squared error (RRMSE) 11](#_Toc164867489)

[3.1.2 Observed heterogeneity 13](#_Toc164867490)

[3.1.3 Correlation 17](#_Toc164867491)

[3.1.4 Estimated RRMSE for different scenarios 17](#_Toc164867492)

[3.2 Existing coverage and gaps in coverage 18](#_Toc164867493)

[3.3 Quality of the admin data and potential overcount 22](#_Toc164867494)

[3.4 Summary and conclusion 25](#_Toc164867495)

[4. Overcount correction 26](#_Toc164867496)

[5. Bias correction 26](#_Toc164867497)

[6. Model selection process 27](#_Toc164867498)

[6.1 The stages of model selection 28](#_Toc164867499)

[7. Household estimation report 32](#_Toc164867500)

[7.1 Collapsing of main household variables 33](#_Toc164867501)

[7.2 Household variable levels before collapsing 34](#_Toc164867502)

[8. Person estimation report 36](#_Toc164867503)

[8.1 Collapsing of main individual variables 38](#_Toc164867504)

[8.2 Individual variable levels before collapsing 40](#_Toc164867505)

[9. Overcount correction report 43](#_Toc164867506)

[10. Bias correction report 43](#_Toc164867507)

[11. Conclusion 43](#_Toc164867508)

[12. Glossary 43](#_Toc164867509)

# Introduction

## *Background*

For over 200 years, Scotland has relied on the census to underpin national and local decision making. It provides anonymous census estimates which offer a highly accurate picture of the number of people and their characteristics (such as age, health, where and how we live etc.). National and local government, the education and academic communities, the third sector, commercial business and others use census information in order to plan and provide their operations efficiently and effectively. The information is particularly important when there is no other reliable source or when the ability to cross-reference or compare characteristics of people or households is required.

The census aims to capture details of the whole population of Scotland. However, it is always expected in any census that some people and households will not submit a return. To mitigate for this we use estimation to adjust for this undercount at both household and individual level, creating estimated population totals for various geographic and demographic groups.

As in previous years our main tool for measuring coverage error is the Census Coverage Survey (CCS). The CCS provides an independent count of a sample of Scotland’s population, measured shortly after the Census. This data is then linked to Census returns, allowing coverage rates to be measured using capture-recapture.

Our plan for coverage estimation, similar to that in 2011, was to create population level estimate for many small groups like age-sex by local authority using separate estimators in a multi-step process based on dual system estimation within each sampled strata, then ratio estimation to extrapolate to non-sampled areas across broad geographic regions. Finally a small area estimation model would be used to break these estimates down over smaller geographic regions. Over coverage would then be measured separately and the results of these two models are combined.

## *Lower than expected response rates*

Scotland’s Census 2022 had a target person response rate of 94% nationally and management information on household return rates were used as a proxy during live operations to measure progress. The final household return rate was 89.2%. The final response rate for the CCS was 57.8% which, though lower than anticipated, was similar to that achieved by ONS for their CCS. It should be noted that the response rate and return rate measure slightly different things, for more information see the glossary in section 12 or the published quality report[[1]](#footnote-1).

Given these low response rates, we were concerned that our planned approach to population estimation would result in lower quality population estimates. During the live census collect operation in order to get an early idea of the potential impact on the quality of the estimates we looked at the width of the confidence intervals around our census population estimates by running simulations of a number of different response rate scenarios to determine the likely change in the confidence interval % (CI%) scenarios including a 89% return rate scenario. The results of which are shown in figure 1 below. For our census RR of 89% with a 70% CCS response scenario, we predicted an error of 0.447%, over our target KPI. Given that the CCS response rate was considerably lower than that, this was likely to be an under estimate of error. This also makes the assumption that there is no dependence between the CCS and census. However it is likely that dependence would increase given lower response rates, causing bias.

Figure 1: Simulations of CI% resulting from various response rate scenarios

An International Steering Group (ISG) was established in response to the challenges faced during the collection phase and NRS worked with these experts as we developed the methods discussed in this paper. On advice from ISG, in order to mitigate the risk of lower accuracy in our population estimates, we made a number of changes to our methodology, which we documented in June 2023 [[2]](#footnote-2). This paper gives further details about the following two changes:

* Change from stratified multistage DSE (more in section 2) to a logistic regression model for under coverage estimation, to allow greater flexibility and statistical efficiency.
* Use of administrative data to supplement estimation, specifically to enhance the Census Coverage Survey (covered in section 3).

Finally the paper will report on the outcomes of the live running of estimation (section 6, 7 and 8).

# Change of DSE model

## Under coverage in 2011

As mentioned previously, our planned approach to coverage estimation, and the method used in 2011, was a three stage process based on stratified DSE at a low level.

### Stage 1 – Dual-system estimation

After matching between the census and the CCS, a 2×2 table of counts of people or households can be derived. This is shown in Table 1.

Table 1: 2×2 Table of Counts of People (or households)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | *CCS* | | |
| Counted | Missed | **Total** |
| *Census* | Counted | n11 | n10 | **n1+** |
| Missed | n01 | n00 | **n0+** |
| **Total** | **n+1** | **n+0** | **n++** |

This output from the matching process was used to estimate the undercount, represented by n00 in the table above, for each of the sampled clusters. Given the assumptions, dual-system estimation (DSE) combines those people counted in the census and/or CCS and estimates those people missed by using the following formula to calculate the total population:

This was calculated separately for each of the strata (cluster and: sex and five-year age group for persons; tenure and, where applicable, type of landlord for households).

### Stage 2 – Estimation area population estimates

The DSEs calculated in stage 1 only produced population estimates for the areas in the CCS sample. In the second stage a statistical model was used to obtain estimates for the non-sampled areas. Of the 32 LAs in Scotland, only Glasgow had sufficient CCS sample to allow a direct LA-level estimate with an acceptable level of precision. The remaining LAs were grouped together at the estimation stage into Estimation Areas (EAs)

Within each EA, a simple ratio estimator (which uses a straight line of best fit through the origin) was used to estimate the relationship in the sample between the census count and the DSE. The weight was calculated separately for each age-sex group within each Hard-to-Count index (HtC) [[3]](#footnote-3) stratum and applied to the census totals for a true population estimate:

A similar approach was used for households, in this case stratifying by tenure.

### Stage 3 – Synthetic LA estimation

Due to the previously mentioned need to calculate the first two stages at EA level, a further stage was needed to calculate estimates of the population by age and sex for persons and household size for households in each LA.

A synthetic estimator was applied to produce LA-level population estimates that have lower variances than those that would be produced by just using the sample specify to each LA. The technique used information from the whole EA to model coverage within the LAs. It made the assumption that coverage levels across the LAs are the same within each HtC and age-sex stratum. The resulting population (and household) estimates will then be calibrated to the EA estimates to ensure consistency, and their accuracy was also calculated to provide confidence intervals around the LA population estimates.

## Logistic regression approach

Instead of this stratified DSE approach, based on advice from our International Steering Group, we decided to use an estimation method based on logistic regression, the same approach Office for National Statistics (ONS) used for their 2021 Census, to arrive at coverage probabilities for a range of demographic groups which would have greater statistical efficiency and flexibility to deal with the lower than expected response. This approach has a number of advantages over the stratified DSE approach and we are grateful to ONS for sharing their code and expertise with us.

By utilising a single regression model (or a small number of regional models) rather than multiple stratified dual system estimation calculations, the increased statistical power allows for a larger set of characteristics to be modelled. This approach essentially integrates the dual-system, Ratio and Local Authority estimation parts of the 2011 methods into a single modelling step followed by a derivation and application of non-response weights.

This approach continues to use the CCS, which is linked to census data giving a list of records in the census alone, CCS alone or in both. A logistic regression model can then be fit to arrive at the probability, y, of a particular individual (or household) record, i, having a census return, given a set of predictors, k:

Once the model is fitted it can be applied to each census record to give the estimated response probability. The inverse of this gives a non-response weight for that record. Rearranging the formula above this gives:

These weights can then be summed for any aggregation of the population, for example the traditional DSE strata (age, sex, LA, HtC), to give an estimated population size.

We expected regional differences in response probabilities. There were two ways to model this: treating the regional variable as a fixed effect or using a mixed effects model with a regional variable as a random effect. A random effects model is more statistically efficient because it allows for variation across regions without estimating a separate effect for each one and is increasingly common practice when modelling data that is structured within multiple levels, but it may not capture all sources of regional differences. To address this, we included Local Authorities (LAs) as a random effect while also grouping LAs into broader regions as fixed effects. These regional groupings were chosen during model selection to best fit the data.

## Advantages of change in model approach

The logistic regression modelling approach adopted here has a number of advantages when compared to the previous approach we had planned for Scotland’s Census 2022. Many of these were particularly beneficial with a lower than expected response rate.

### Reduction in heterogeneity bias

The use of a single model across the nation, rather than separately within regional estimation areas and demographic strata, gives greater statistical power to include more predictors. This means that we are better able to control for other demographic factors that influence return rates beyond age, sex and HtC, thereby reducing bias due to heterogeneous response probabilities across the population in both census and CCS. ONS research suggests that this could reduce absolute relative bias by 0.2% (our KPI is below 0.5%)[[4]](#footnote-4). This is of particular importance in a low response rate scenario, as the concern is that some of the populations which drive the low response rate in the census will also respond poorly to the CCS, meaning a loss of statistical power under the planned traditional DSE model. The use of a single model allows us to better identify and control for these segments of the population, and thus correct for them without creating undercount.

### Reduction in LA level bias due to synthetic estimator

Another benefit to using a combined model is that we would be less reliant on achieving homogenous response rates across estimation areas. This goes beyond the previous point about heterogonous response probabilities, as, even given 100% coverage in the CCS, the use of a synthetic estimator for LA population counts can cause bias if there is variation in census response rates between regions. This kind of bias, with some LAs population counts being biased upwards and some downwards, can be particularly well dealt with by mixed effects model with LA as a random effect and use of region (grouped LAs) as a fixed effect. Given the lower than expected response rates to Scotland’s Census 2022, there is the potential for greater variation between LA response rates.

### Lower estimate random error

As the proposed approach pools statistical power across the nation it should reduce error in the population estimates due to sampling error. This is because by simultaneously modelling demographic predictors across regions should effectively mean there is more data available to estimate predictor coefficients, counteracting any outliers in the sampling distribution. By including more predictors than used in the previously planned DSE strata, this should bring down error in population estimates. NRS simulations (carried out before implementation) have shown that in a lower response rate scenario (89% nationally) confidence interval may be around +- 0.45%[[5]](#footnote-5), above our KPI of +- 0.4%, making such an improvement particularly desirable.

### Possibility of using operational data in model

Unlike in the current approach using DSE, a logistic regression allows us to use operational data (i.e. return rates) in modelling the probability of non-response. This allows another source of information to be used in calculating the level of undercount region by region, improving the quality of the population estimates

### Methodologically streamlined – preliminary low level estimates early

The use of a single model to arrive at non-response weights makes the process more streamlined. There is no need for multiple steps to arrive at LA level estimates, and the weights can be used to arrive at population estimates for other geographic or demographic breakdowns. This allowed us to start preliminary validation of these low level geographic areas and non-estimation variables ahead of estimation completion, allowing feedback and ahead adjustment and post-adjustment imputation. This allowed us to deliver first outputs in a timely manner with less risk to quality.

## Challenges to implementation

The main challenge to the implementation of the logistic regression approach was that the modelling process requires more analyst involvement and decision-making about model construction. Considering the number of possible predictors, and the possible interactions between them, there are a huge number of possible models to choose from. Care had to be taken to ensure that model selection process was systematic and data driven, as poor model specification could lead to bias. The process and steps we took to mitigate this risk are discussed later in the report section of this paper. In addition, decisions about modelling were taken under the oversight both our internal board and the external International Steering Group.

# Use of administrative data to enhance population estimation (under coverage model)

As discussed, the lower than expected return rate meant that we had to consider a number of options to mitigate associated increases to estimate random error and bias. One of these was the use of an administrative data spine to boost numbers in one of the two lists (census and CCS) being fed into our undercoverage model.

The procedure for enhancing with admin data was first to derive an admin data spine by combining NHSCR, electoral roll, School Pupil Census, Health Activity, Higher Education Statistics Agency (HESA) and NRS Vital Events (births, deaths and marriages). These records were first linked to the census and CCS in order to select only records that are not already part of these lists. So when adding to the census, admin records already in census would not be added, and in the same way records already in the CCS would be excluded when adding admin data to that list. Once admin data has been added to one of the two lists, it was then linked more thoroughly to the other list to enable estimation (which requires a count of those in either list only and in both).

We had to take care to add only records that represent real people still in the population on census day. We did this by using inclusion criteria based on the strength of evidence which is a rating of how many of the admin data sources it appears on. Admin data records that have corresponding records on a number of sources are more likely to be people still in the population. We also implemented rules so that people were not added into existing houses, and not added in a situation in which the created house had more than 6 members. Finally, once a list has been enhanced, it would be fed into estimation in the usual way.

The decision of which list to enhance with administrative data depended on the overlap between the admin data spine and the two lists. For this reason, a decision on which list to enhance was postponed until the administrative data had been compiled, linked to the census and CCS, and made available for analysis.

Once these processes were complete we could answer the question of which list enhancement will give optimal population estimates.

This section summarises this analysis leading to our decision to supplement the CCS list with administrative data. There are three areas of analysis that were considered in this decision:

* What impact does enhancing the two lists have on estimate prediction error, and which list enhancement produces the least error
* Were there any geographical areas which have significant gaps in coverage in either census or CCS, which might have suggested we need to enhance the list with admin data
* Were there any quality concerns around the administrative data, particularly concerning overcount or bias

A draft of these proposals and associated analysis were taken to the International Steering Group, who concurred with our recommendations.

## Population precision and the impact of admin data enhancement on error

### Derivation of relative route mean squared error (RRMSE)

It is possible to calculate the likely gains in estimate precision from enhancing either the census or the CCS.

From a simple two-way table defining the joint relationship between census response and survey response; we can define to be the marginal probability of being in the census, to be the marginal probability of being in the Census Coverage Survey, and as the probability of appearing in the census and the CCS. The actual counts are then just scaled by the population total , which for simplicity we will set as 1. It then follows that an estimate of applying dual-system estimation is

and following from Wolter (1986) its variance is approximately

We can immediately see how increasing either or through enhancing the response with records from an administrative list will reduce the variance coming from the dual-system component of estimation. However, given the potential for much greater increase to due to the relatively lower initial survey response, it implies the greater reduction will be from enhancing the survey.

Once the dual-system component has created estimates of the true population counts for the sample areas indexed , we need to ratio up from the sample areas to the population. We use an estimator of the form

where is the census count for an area , is the CCS sample of areas, and is the total census count. Within our simple model, where we are looking at the probability structure underlying a sample area by setting to one, this simplifies to

From classical sampling, the variance of a ratio type estimator (with ) is of the form

where is the scale of the sampling variance for a standard estimator of the total of or , and is the underlying correlation between the values and . Therefore, in our case we need to scale our dual-system component by

leading to the relative root mean square error given by

where we continue with set at one.

This is now where enhancing the census comes into play more directly. This will increase but not , which moves the coverage ratio closer to one, reducing variance from the ratio component. Based on the working assumption that the administrative record list is created independent of census and survey response, enhancing the survey does not, on average, change the ratio as records are added to both in similar proportions in that scenario. In addition, enhancing the census response might be expected to improve the correlation by smoothing out variation in the census returns across local areas. This will not happen with enhancing the survey response, although stabilising the dual-system estimates for local areas will (likely) improve the observed correlation in the sample areas to some extent.

So far, the discussion is framed in the context of the 2011 approach to coverage estimation, see Brown *et al* (2019) for the general implementation; but our current approach is to use a logistic regression approach (LREM). This fits a model to the probability at the individual level, allowing for borrowing strength across areas while embedding more detailed effects such as an individual’s household tenure or their ethnicity. In this case, it still follows that enhancing the survey will boost the amount of data in the modelling; while enhancing the census will likely improve the prediction element. However, we will be estimating a (very) common event, enhanced census coverage, using the relatively low survey response. We note that this is the exact scenario that ONS faced post-2021 with the exception that it was driven by a high initial response to census, not an improved response created by enhancing with administrative records.

Based on the above, the question of which list (census or CCS) to enhance to bring about the largest reduction in error comes down to the balance between the gain of the model precision on one hand (through enhancing the CCS) which can easily be calculated from coverage rates; and the enhancement of prediction from the model (through enhancing the census) which depends on the reduction in heterogeneity in the census when admin data is added. Analysis of this, together with the observed correlation ( above) between census counts and population estimates, is covered in the next section.

### Observed heterogeneity

To analyse how the correlation between census returns and the underlying population changes when admin data is used to enhance the census, we looked at both the patterns of both return rates and estimated response rates before and after introducing admin data.

*Return rates*

For return rates, we used household return rates and compared these to estimated household return rates when admin data was added. These estimates were calculated by taking the census return rates for each LA by HTC strata and scaling them up in line with the increase in the individual response rate for that strata when admin data was added.

Looking at figure 2 below, while return rates increase when admin data is added, and there is some smoothing across strata, this smoothing effect is fairly minimal, with the original patterns of response more or less maintained.

Graphical user interface, application

Description automatically generated

Figure 2: Household return rates and estimated return rates with admin data

Chart

Description automatically generated with medium confidence

Figure 3: Estimated census individual response rates and individual response rates after adding admin data

*Response rates*

Response rate for the census in a particular area can be estimated as approximating to the CCS linkage rate for that area. This is because, by the above equation the true population count is equivalent to the inverse of this rate multiplied by the census count for that strata. Dividing the census count by this (as you would to arrive at response rates), clearly brings us back to the linkage rate.

These response rates are plotted in figure 3 above, for each LA and HTC. We can see that there is again little smoothing when admin data is added, although the derived response rates do increase.

We also looked at the distribution of these response rates across planning areas for specific LAs. While we found some reduction in the variation across planning areas when admin data is added, it was not particularly large, and in some areas it in fact increases.

### Correlation

In order to directly estimate and how it changes when admin data is introduced we used linkage rates to derive population estimates (as above) by post code.

We then calculated the partial correlation between estimates and census values (controlling for effect of LA and HTC). These were:

Census only = 0.9862176

Census + admin = 0.9885093

### Estimated RRMSE for different scenarios

Using the calculated correlations together with census and CCS return rates and estimated admin coverage rates (assuming conservative rules around the standard of evidence required to be taken as representing a real person), the RRMSE can be estimated using the formula above (see table 1 above). This gives 2.73% for adding the admin data into the census, and 3.13% for adding it into the CCS. This would suggest that adding the data to the census would give the best gain in terms of estimate precision.

The correlations derived above seem to be surprisingly high. They would suggest that almost all variation in Census response was between LA and HTC, with very little being driven by respondent demographics. This seems unlikely and contrary to findings elsewhere. This is likely due to the fact that by performing this analysis at the household level and aggregating up to post codes, we have removed the individual variation in response.

To account for the possibility that the correlations suggest a more homogeneous response rate than was in fact the case, we also looked at more conservative correlation value, reduced by .05. In this scenario the model with admin data added to the CCS performs better with a RRMSE of 5.79% compared to 6.16% when the admin data is added to the census. It seems then that as the heterogeneity increases, a CCS enhanced with admin data is better able to cope with it than the census enhanced with admin data. This suggests that supplementing the CCS is the more conservative strategy, as if response rates are actually as homogenous as the initial analysis suggests, modelling it should be relatively straightforward anyway.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Admin coverage | Bias (odds ratio) | Census + admin correlation | Census only correlation | RRMSE for census + admin vs CCS | RRMSE for census vs CCS + admin | Bias: census + admin vs CCS | Bias: census vs CCS + admin |
| 0.66 | 0 | 0.9885093 | 0.9862176 | 2.73% | 3.13% | 0 | 0 |
| 0.66 | 0 | 0.9385093 | 0.9362176 | 6.16% | 5.79% | 0 | 0 |
| 0.66 | 1.4 | 0.9862176 | 0.9862176 | 2.75% | 3.11% | -0.50% | -0.36% |
| 0.66 | 0.7 | 0.9862176 | 0.9862176 | 2.81% | 3.19% | 0.52% | 0.37% |

Finally, it may be that there is some dependence bias between the census and CCS which would lead the estimate to be biased. These scenarios were also compared, and it seems that an admin data enhanced CCS is better able to deal with dependence bias than an admin data enhanced census (at least when CCS response rates are lower than census response rates, as is the case for us).

Table 1: Table of RRMSE and bias estimates for different scenarios

## Existing coverage and gaps in coverage

Another point we considered in the decision of which list to enhance is whether there are any serious current gaps in coverage in the two lists that would necessitate their enhancement with admin data. As well as leading to higher error at a national level, such heterogeneity could introduce biases, with some areas receiving either to large or too small a coverage correction. As we saw in the previous section, although there are variations in census coverage, there are no areas in which the coverage levels could be said to be critically low. The lowest returning LA-HTC strata is HTC 4 in Aberdeenshire with 68.5% and the lowest returning planning area is in Glasgow HTC 4 with 52.9%.

The other question is whether there were any areas with particularly low CCS coverage that would have made enhancement of that list preferable. As you can see in figure 4 bellow, as the CCS coverage rates are generally lower, there are more areas with lower return rates for the CCS (the lowest response rate area is in North Lanarkshire HTC 5 with a rate of 19.1%). When admin data is added these discrepancies are smoothed out, with the lowest return rate being 38.3% in North Lanarkshire HTC 5 (an almost 20pp increase).

Chart

Description automatically generated

Figure 4: Heatmap showing return rates for the CCS by LA-HTC strata, before and after adding admin data

Again we also looked at the distribution of these return rates across planning areas for specific LAs . We found that that in some cases the distribution before adding admin data was quite wide. However, when the admin data was added the distributions for the most part show a decrease, suggesting that the estimates from this data would be more stable, allowing a precise estimate of undercount even in lower geographical areas, adding to the case for enhancing the CCS.

## Quality of the admin data and potential overcount

The final question was around the quality of the admin data and how that contributed to the decision of whether to add it to the census or the CCS. One of the key questions was whether the age-sex distributions of the admin data are similar to the general population, or whether certain demographics were over-represented. To examine this, we compared the census distribution to the admin distribution. We also looked at a census and admin combined figure, a population estimate using dual system estimation with admin and census, and the census day estimate derived from mid-year estimates.

We can see this for males and females in figures 5 and 6 below. In general the coverage patterns of the census and admin data (and their combination) follow the coverage patterns in the census day estimate, showing that they are broadly representative of the wider population. The estimate from admin data is also broadly in line with the census day estimate, although it is slightly higher.

There are a few areas where things deviated. There was a dip in the distribution for the census day estimate around age 20, this was not as deep in the admin data or admin data estimate (especially for females). There was also an peak in the census day estimate around age 30, which was not represented fully in the admin data or admin data estimate. However, given the census seems to corroborate the admin data, it is reasonably plausible that both of these were issues in the census day estimate rather than the admin data and census. A more serious issue was the slightly elevated admin data + census counts for ages over 50. This may be due to overcount in the admin data in these age groups, and it is worth re-evaluating our inclusion criteria for admin data in this age range specifically.

Chart, line chart

Description automatically generated

Figure 5: Age distributions of admin data and comparators: males

Chart, line chart

Description automatically generated

Figure 6: Age distributions of admin data and comparators: females

A similar population estimate was created using the administrative data and CCS in a dual system estimation. The results of this are shown in figures 7 and 8 for males and females. We can see that this produces a notable peak in the distribution at age 20 to 30, which is considerably higher (around 1200 for females) than the estimate from census in CCS areas only (around 840, not shown here). It is also considerably higher than the peak seen in the census day estimate. The reason for this appears to be that our linking procedure was not able to detect some matches between admin data and the CCS, due to the non-standardised address information on the survey. This means that more records appear to be only on one list in the dual system estimation (admin or CCS), causing an increased population estimate.

Chart, line chart, histogram

Description automatically generated

Figure 7: Age distribution of admin data, CCS and dual system estimate: males

Chart, line chart

Description automatically generated

Figure 8: Age distribution of admin data, CCS and dual system estimate: females

This finding enhanced the case for adding the admin data to the CCS, as the estimate from CCS and admin mirrors the scenario where the Census list is enhanced (as in that case the estimate would be Census + admin vs CCS, similar to admin vs CCS shown here). Additionally as the CCS is a smaller sample we were able to conduct more thorough linking between it and the administrative data in CCS areas, as well as between that admin data and the Census. If we had added the admin data to the Census, we would have had to link the entire admin spine to the census, which would have limited our ability to do the highest quality of linking. In addition, we were able to adopt, very conservative, criteria for the deduplication linking to CCS, to mitigate for the potential overcount. We would not have been able to do this between admin and CCS for estimation (rather than deplication) where we need to be correct, and cannot err on the side of caution. Therefore by adding the admin data to the CCS then the linking for estimation is between this the combined CCS and admin dataset and the census, where we think linking accuracy is good, and the linking to the CCS is deduplication, where we can be conservative.

## **Decision: enhance the coverage survey with administrative data**

Our three criteria to determine whether to enhance the census or CCS with administrative data were:

* The impact on bias and prediction error;
* the potential for enhancement to smooth out differences in coverage in either list;
* and the quality of the administrative data, and considerations about our ability to deal with duplication via linking.

We found that enhancing the census or the CCS with administrative data each reduces the overall error. If the correlation in coverage rates is as high as initial calculations suggest, adding administrative data to the census yields a slightly lower relative error. However, if coverage is more heterogeneous than first assumed, adding administrative data to the CCS performs better, which seem plausible as correlations did not take into account within post code. In addition, the CCS can handle dependence much bias more robustly.

We compared the heterogeneity in return and response rates for both the census and CCS before and after adding administrative data. This revealed that the CCS showed the most improvement and had the greater initial heterogeneity. The census, on the other hand, did not show a large amount of heterogeneity before adding administrative data, likely due the strategic priority in collection to minimise return rate variations.

Overall, the findings indicate that the census and administrative data estimate (which is a proxy for CCS enhancement), as well as their combination, broadly reflect the age and sex distribution of the wider population, although the administrative data tends to slightly overestimate in certain age groups. On the other hand the dual system estimation using admin and CCS data (a proxy for enhancing the census with administrative data) produced an exaggerated peak in the 20–30 age range, likely due to matching issues caused by non-standardised address information. As enhancing the CCS allows for more accurate and conservative linking between the combined CCS-admin dataset and the census as well as deduplication of the CCS and admin data, these results bolster the case for adding the admin data to the CCS.

Therefore, on all three criteria our analysis showed that enhancing the CCS was preferable to enhancing the census, so we took the decision to enhance the CCS by adding in administrative records for individuals who had not submitted a return.

# Overcount correction

The two main parts of overcount that we are able to measure and correct for are duplication, where someone has responded to the census in two different places, and misplacement, where someone has responded once to the census but not at their usual address. Duplication within a household or postcode would be removed in earlier processes such as Resolve Multiple Responses (RMR), but some still remained in the data when duplications are across different postcodes.

Census to census linking was used to identify potential duplicates within the census data, with further linking to admin data used to assign probabilities for the record representing a unique person in the dataset. The methodology for this linking is detailed in [PMP015: Census to census linking and overcount correction | Scotland's Census (scotlandscensus.gov.uk)](https://www.scotlandscensus.gov.uk/documents/pmp015-census-to-census-linking-and-overcount-correction/). This gave an overcount duplication probability for every record.

Census to CCS linking was used to identify cases where someone was in a different place between the census and CCS, though can only tell us about people who live in CCS sample areas. Counting only the records where the respondent has not indicated that they moved since census day as misplaced in the census, we can group the records in to age group strata and produce overcount misplacement probabilities to be applied across the non-CCS sample area records.

The duplication and misplacement probabilities were then combined to produce an overall overcount weight for each record, which would take a value of less than 1. This was combined with the undercoverage weight from the logistic regression model for each record to get an overcount-corrected estimation weight.

# Bias correction

In the ideal situation for Estimation, the probability of a household responding to the CCS is independent of whether or not the household responded to the census. As this assumption does not entirely hold in practice, the estimates can end up with bias due to the dependence between the two. As it is more likely that a census non-respondent would also not respond to the CCS, for example where people are sceptical of answering any surveys, the dependence historically has given rise to bias that results in an underestimate of the number of households.

In order to determine if there was dependence bias, and if so how much, we compared our census household estimate to the rolled forward NRS Household Estimates, and to an alternative household estimate (AHE) that we constructed using census operational data and administrative sources. These sources were particularly useful as they utilised council tax data to identify households, which was not part of our administrative data for CCS enhancement.

In the case that bias was detected in a Local Authority, we used these alternative comparator sources to set a target number of households to add and counteract the bias. The households were added to maintain the responding distribution of household sizes, and were propagated to the person level estimates to match the number of individuals needed for each household size

# Model selection process

As discussed, we modelled under-coverage in the census by linking it to CCS survey data enhanced with administrative data in sampled areas. This created a dataset of each survey response or administrative record, their recorded characteristics, and whether they responded to the census or not. We used this to build models of individual and household level non-response.

Logistic regression was the algorithm used to train the models, where the goal was to predict the probability of a person or household responding to the census based on a number of characteristics. For households these included tenure, household size, household structure etc., and for persons these included things such as age-sex, and ethnicity. This was an extension of the dual-system estimation approach used in the 2011 Census.

The model selection process was created to systematically evaluate and identify the best-performing model. The performance of each model was evaluated using appropriate metrics and the best fitting model was then selected and deployed on the census response data to predict the probabilities of each record responding to the census. These (undercoverage) probabilities were then transformed into weights which up-weight the census data. These could be summed up to give the coverage-adjusted estimates. This full process was done separately for persons and households.

As stated above, the training data used to build the models was comprised of data from the census and the (administrative data supplemented) CCS. The CCS was a 1.75% sample of postcodes across Scotland, which meant roughly 45,000 households were included in the sample. However, the population used for modelling was smaller because of non-response to the CCS. A large proportion of this non-response was resolved by supplementing the CCS with the administrative data.

The model selection process included the following steps (carried out for both persons and households separately):

1. Univariate and bivariate analysis of explanatory variables and their interactions, at their collapsed levels. This was used to understand which levels of variables (or levels between interactions) had low counts to then refine collapsing.
2. Variables that were the basis for census outputs were pre-selected to be included in all versions of the model.
3. Purposeful selection was carried out to run models whose predictors were made up of different combinations of main effects (variables) being considered for inclusion in the model. Diagnostics for these models then informed which variables should be kept as main effects.
4. Testing was done for effective interactions between the main effects variables to be included in the model.
5. Stepwise selection was then used to select main effects and interactions. This was done within K-fold cross validation (using five folds). The main metrics used to make decisions about which model to choose were the prediction errors and goodness of fit values estimated from cross-validation.

This process was applied to the two modelling exercises required for coverage estimation which were:

* Household population undercoverage of individuals
* Household population undercoverage of households

## The stages of model selection

The goal of the model selection approach was to select the best possible model, within a set timeframe, this allowed the timeliness quality dimension to be considered as to how quickly census results could be made available to our data users. This approach followed a set of high level principles, including:

* Main variables that would be the basis for census outputs (such as age-sex) were pre-selected to always be included in any version of the model tested.
* Standard descriptive statistics including univariate and bivariate analysis of explanatory variables and their interactions were used to check for numerical issues, at their collapsed levels. This was used to understand which levels of variables (or levels between interactions) had low counts to then refine collapsing and follow good modelling practice.
* First order (two-variable) interactions were analysed by iteratively adding interaction terms one at a time to the model that already included the chosen main effects, and then assessing how each interaction term contributes to the overall model fit.
* To improve model stability, stepwise selection was performed within cross-validation to select the main effects and interactions to be included. This is where second and third order interactions were selected and tested.
* When fitting the models the principle of marginality was maintained by only allowing variables that were included in the model as main effects to be considered for interactions. This hierarchical structure was maintained for higher order interactions.
* The smallest prediction error and the goodness of fit were used to select the chosen models; the variance of coverage weights was not used as a diagnostic.
* To optimise model performance and assess the robustness of the model, multiple tuning parameters were tested to ensure the model performed well across these different parameter settings.

The model selection approach consisted of seven stages. Each stage was carried out sequentially as the outputs from the previous stage were required for the next stage. However, some stages would go through a number of iterations before moving onto the next stage. The process is illustrated in Figure 9.

A diagram of a process

Description automatically generated

Figure 9: The stages of the model selection process

**Stage 1: data preparation**

This stage selected and filtered the data for the specified model (individual or household).

**Stage 2: collapsing and initial data analysis**

In this stage variables could be renamed, transformed and collapsed into levels specified.

This is also where univariate and bivariate descriptive statistics were produced to inform the refining of collapsing as well as get an initial idea of any variables that were likely to provide value to the model.

**Stage 3: purposeful main effect selection**

In this stage the main variables were forced into the model and purposeful main effects selection was carried out.

Variables were iteratively added to and removed from the model to understand their impact on the BIC (Bayesian Information Criterion) score which is an estimator of prediction error that penalises models that use more parameters.

**Stage 4: initial interaction analysis**

Here, each single interaction term was added to the selected main effects model and metrics were produced on the impact on the model fit when adding this interaction. This was done separately for each possible interaction term.

Likelihood ratio tests were used to assess the goodness of fit between the main effects only model and a model with the main effects as well as the interaction term being considered.

In this stage there were also checks for numerical issues such as quasi-complete separation where the outcome variable separates a predictor variable or combination of predictor variables almost completely. Singularity of the covariance matrix was checked for which would occur if there were predictor variables that were perfectly collinear. This can occur when there are fewer observations than variables or when one or more variables are constant or nearly constant across all observations.

**Stage 5: cross validation**

In this stage K-fold cross validation (using five folds) was used to assess candidate models. The main metrics used to make decisions about which model to choose were the prediction errors and goodness of fit values.

Stepwise selection was performed within the cross-validation to select the main effects and interactions to be included. This is where second and third order interactions were selected and tested.

Five-fold cross validation is used to estimate the prediction error of models fit to the CCS data. To carry out this process, the CCS data is divided into five distinct parts with the number of records in each kept as equal as possible. The forward stepwise model selection method is then run five times, using each possible four-fold combination once as the training data before testing prediction error using the remaining part in each case. Different modelling branches and starting main effects/interactions can then be evaluated by comparing mean prediction error which is estimated by averaging the results from scoring on each fold.

From the candidate models, those with the smallest prediction error for logistic and mixed effects logistic models were fit to the full modelling dataset.

**Stage 6: goodness of fit and diagnostics**

Then, the best performing models were checked for any issues.

**Stage 7: variance estimation**

In this stage a bootstrap approach was used to estimate the variance for the undercoverage error-corrected population size estimates. This variance estimation is different from the actual variance estimation used post-model selection to get the population estimates. The one in this stage used a smaller number of replicates and didn’t incorporate any adjustments.

Variance estimation was carried out by producing bootstrap versions of the CCS modelling data and then producing population estimates for each bootstrap. These population estimates are created by scoring the census data using the CCS model coefficients, recalculated when applied to each bootstrap dataset, to calculate the probability of a census response for each record. The reciprocals of these probabilities were then taken to get coverage weights for each record which were then totalled. Confidence intervals and the variance of the population estimates are then calculated for each local authority.

These stages were repeated iteratively many times to explore different collapsing’s, approaches and ideas, before arriving at the final model selected.

# Household estimation report

Following the model selection approach outlined in Section 2, the final household model was as follows.

The main effects included in the model were:

* Accommodation type
* Hard to count index
* Household ethnicity
* Household size
* Household structure
* Tenure
* Region
* Return rate.

The two-factor interactions included in the model were:

* Accommodation type by Household structure
* Accommodation type by Hard to count index
* Accommodation type by Region
* Hard to count index by Region
* Household ethnicity by Household structure
* Household ethnicity by Household size
* Tenure by Household size.

## Household data preparation

The input data that the household models were trained on was the full CCS dataset with the CCS to census links. These links provided the response variable for the logistic regression model which is whether a household responded to the census or not (i.e. the CCS household has a link to a census household).

As noted, the approach was taken to enhance the CCS with administrative data. This process required preparation of the administrative dataset before then creating households from these individual level records. To ensure only records that represented real people still in the population on census day were added, restrictions were placed on the number of records included based on the strength of evidence. The records belonging to the Core strength of evidence group were selected. Rules were implemented so that people were not added into existing houses, and not added in a situation in which the created house had more than 6 members. These individual records were then grouped into households according to their unique address identifiers (UPRN).

A slightly broader inclusion criteria was used in the event that a grouping of individual records resulted in a minor only household. This happened when the minors within a household all existed within the Core group but their parents were in the Standard group of records. The rule was applied that when this occurred, the parents from the Standard group were included in the dataset to resolve the minor only household.

These administrative household were then linked to the census, before being added into the input dataset to supplement the CCS. Some derived variables were created within this input dataset to create features which could provide additional value to the models being created. These included features such as household ethnicity and household structure.

This input dataset was then fed into the model selection system. The final model selected was then applied to the complete census household dataset. This creates the population estimates by scoring the census data using the CCS model coefficients to calculate the probability of a census response for each household record.

## Collapsing of main household variables

As outlined in the model selection approach, collapsing of the main variables was done to enable numerical issues to be resolved. This allowed for interactions to be included in the model that provided value and were representative of the characteristics of those responding to the census and the CCS that without collapsing wouldn’t have been suitable for inclusion in the model.

The collapsed levels for household size were:

* One
* Two
* Three
* Four
* Five plus

The collapsed levels for tenure were:

* Owns outright, with shared equity or part owns and part rents
* Owns with a mortgage
* Rents from council or housing association
* Rents from other (including private landlord/ letting agency) or rent free

The collapsed levels for accommodation type were:

* House
* Flat or temporary or mobile

The collapsed levels for household ethnicity were:

* White
* Mixed, Asian, African, Black or Caribbean, other, and different ethnicities within household

The collapsed levels for region were:

* Dumfries & Galloway, Na h-Eileanan Siar, Highlands, Orkney Islands, Shetland Islands, Aberdeenshire, and East Dunbartonshire
* Edinburgh, and Glasgow
* Aberdeen City, and Dundee City
* East Renfrewshire, Moray, Argyll & Bute, East Ayrshire, Falkirk, Midlothian, and South Lanarkshire
* Clackmannanshire, Inverclyde, North Ayrshire, Renfrewshire, West Dunbartonshire, West Lothian, and North Lanarkshire
* East Lothian, Scottish Borders, South Ayrshire, Stirling, Angus, Fife, and Perth & Kinross

The 24 levels of household structure were collapsed into:

* Single male or female parent, or male or female one person household, aged 16 to 34 years
* Single male parent, or male one person household, aged 35 to 64 years
* Single female parent, or female one person household, aged 35 to 6 years
* Single parent (male or female), or one person household (male or female), aged 65 years and over
* Couple, aged 16 to 34 years, with or without children
* Couple, aged 35 years and over, with or without children
* Couple with children where some children are adults (any age), and related persons not yet categorised aged 35 years and over
* Related persons not yet categorised, aged 16 to 34 years
* Unrelated persons not yet categorised aged 16 years and over, and minor only households

## Household variable levels before collapsing

The household main variables before collapsing with original levels were as follows.

**Accommodation type**

* House
* Flat
* Temporary or mobile

**Hard to count index**

* One
* Two
* Three
* Four
* Five

**Household ethnicity**

The household ethnicity variable is defined as the ethnicity of all individuals within the household, if they all have the same ethnicity.

* African
* Asian
* Black or Caribbean
* Different ethnicities within household
* Mixed
* Other
* White

**Household size**

* One
* Two
* Three
* Four
* Five plus

**Household structure**

The household structure variable is defined by the following broad categories.

* Single adult only
* Single adult with children
* Couple only
* Couple with children
* Couple with children where some children are adults
* Related persons not yet categorised
* Unrelated persons not yet categorised
* Minor only households

Within these broad categories, sub-categories are created where appropriate, stratified by the age and sex of the adults in the household, for a total of 24 categories.

**Tenure**

* Owns outright
* Owns with a mortgage
* Owned with shared equity
* Rents from council or housing association
* Rents from private landlord or letting agency
* Rents from other
* Part owns and part rents
* Rent free from council or housing association
* Rent free from private landlord or letting agency
* Rent free from other

**Region**

* Dumfries and Galloway, Aberdeenshire, East Dunbartonshire, Highlands, Orkney Islands, Shetland Islands and Na h-Eileanan Siar
* Glasgow
* Dundee City and Aberdeen City
* Argyle and Bute, East Renfrewshire, Moray
* Clackmannanshire, Inverclyde, North Ayrshire, North Lanarkshire, Renfrewshire, West Dunbartonshire and West Lothian
* Angus, East Lothian, Fife, Perth and Kinross, Scottish Borders, South Ayrshire and Stirling
* East Ayrshire, Falkirk, Midlothian and South Lanarkshire
* Edinburgh

**Return rate**

Continuous variable.

# Person estimation report

Following the model selection approach outlined in Section 2, the final individual model was as follows.

The main effects included in the model were:

* Return rate
* Activity last week
* Address one year ago
* Born UK
* Household size
* Hard to count index
* Marital status
* Ethnicity
* Full time student
* Tenure
* Sex
* Accommodation type
* Relation to others in household
* Age(banded)
* Region

The two-factor interactions included in the model were:

* Return rate by hard-to-count index
* Household size by hard-to-count index
* Household size by marital status
* Activity last week by fulltime student
* Household size by fulltime student
* Ethnicity by student
* Return rate by tenure
* Activity last week by tenure
* Household size by tenure
* Hard-to-count index by tenure
* Household size by accommodation type
* Hard-to-count index by accommodation type
* Tenure by accommodation type
* Address one year ago by relation to others in household
* Tenure by relation to others in household
* Sex by relation to others in household
* Accommodation type by region
* Student by age band 9-17
* Tenure by age band 9-17
* Return rate by age band 18-22
* Address one year ago by age band 18-22
* Tenure by age band 18-22
* Address one year ago by age band 23-26
* Tenure by age band 54-58
* Born UK by 59-64
* Tenure by age band 59-64
* Born UK by 65-71
* Tenure by age band 65-71
* Household size by 72-78
* Tenure by age band 72-78
* Activity last week by age band 79+
* Born UK by age band 79+
* Tenure by age band 79+

## Person data preparation

The input data that the person models were trained on was the full CCS dataset with the CCS to census links. These links provided the response variable for the logistic regression model which is whether an individual responded to the census or not (i.e. the CCS household has a link to a census household).

As noted, the approach was taken to enhance the CCS with administrative data. This process required preparation of the administrative dataset. To ensure only records that represented real people still in the population on census day were added, restrictions were placed on the number of records included based on the strength of evidence. The records belonging to the Core strength of evidence group were selected. Rules were implemented so that people were not added into existing houses, and not added in a situation in which the created house had more than 6 members.

A slightly broader inclusion criteria was used in the event that a grouping of individual records resulted in a minor only household. This happened when the minors within a household all existed within the Core group but their parents were in the Standard group of records. The rule was applied that when this occurred, the parents from the Standard group were included in the dataset to resolve the minor only household.

These individual level administrative records were linked to the census, before being added into the input dataset to supplement the CCS.

This input dataset was then fed into the model selection system. The final model selected was then applied to the complete census individual dataset. This creates the population estimates by scoring the census data using the CCS model coefficients to calculate the probability of a census response for each individual record.

## Collapsing of main individual variables

As outlined in the model selection approach, collapsing of the main variables was done to enable numerical issues to be resolved. This allowed for interactions to be included in the model that provided value and were representative of the characteristics of those responding to the census and the CCS that without collapsing wouldn’t have been suitable for inclusion in the model.

The collapsed levels for household size were:

* One
* Two
* Three
* Four
* Five plus

The collapsed levels for related to others in household were:

* Related to someone in the household
* Not related to someone in the household

The collapsed levels for address one year ago were:

* Same address as current
* Other address

The collapsed levels for age grouping were:

* 0-2 years old
* 3-8 years old
* 9-17 years old
* 18-22 years old
* 23-26 years old
* 27-30 years old
* 31-34 years old
* 35-39 years old
* 40-43 years old
* 44-48 years old
* 49-53 years old
* 54-58 years old
* 59-64 years old
* 65-71 years old
* 72-78 years old
* 79+ years old

The collapsed levels for fulltime student were:

* Yes, is a fulltime student
* No, is not a fulltime student

The collapsed levels for born in the UK were:

* Born in the UK
* Not born in the UK

The collapsed levels for marital status were:

* Never married or registered in a civil partnership, or under 16
* Married or in a registered civil partnership
* Other/ Married or in a civil partnership but are now separated, legally divorced/dissolved or widowed

The collapsed levels for ethnicity were:

* White
* Not white

The collapsed levels for accommodation type were:

* A whole house or bungalow that is detached, semi-detached or terraced
* A flat, maisonette or apartment that is in a tenement, part of a converted or shared house or in a commercial building or mobile/ temporary structure

The collapsed levels for tenure were:

* Owns with a mortgage, loan or shared equity or part rents and part owns
* Owns outright
* Rents from council or housing association
* Rents from private landlord, letting agency, other or lives rent free with landlord: council, letting agency, private landlord or other

The collapsed levels for activity last week were:

* Employed
* Looking/ waiting for work or long term sick/ disabled
* Student, retired, Looking after family, other

The collapsed levels for region were:

* Dumfries and Galloway, Aberdeenshire, East Dunbartonshire, Highlands, Orkney Islands, Shetland Islands and Na h-Eileanan Siar
* Glasgow and Edinburgh
* Dundee City and Aberdeen City
* Argyle and Bute, East Renfrewshire, Moray, East Ayrshire, Falkirk, Midlothian and South Lanarkshire
* Clackmannanshire, Inverclyde, North Ayrshire, North Lanarkshire, Renfrewshire, West Dunbartonshire and West Lothian
* Angus, East Lothian, Fife, Perth and Kinross, Scottish Borders, South Ayrshire and Stirling

## Individual variable levels before collapsing

The individual main variables before collapsing with original levels were as follows.

**Household size**

Continuous variable.

**Related to others in household**

* Related to someone in the household
* Not related to someone in the household

**Address one year ago**

* Same address
* Other address UK
* Other address non-UK

**Age**

Continuous variable.

**Fulltime student**

* Yes, is a fulltime student
* No, is not a fulltime student

**Born in the UK**

* Born in the Scotland
* Born in England
* Born in Wales
* Born in Northern Ireland
* Born in Republic of Ireland
* Born elsewhere

**Marital status**

* Never married and never registered in a civil partnership
* Married
* In a registered civil partnership
* Separated but still legally married
* Separated but still legally in a civil partnership
* Divorced
* Formerly in a civil partnership which is now legally dissolved
* Widowed
* Surviving partner from a civil partnership

**Ethnicity**

* White
* Mixed or multiple ethnic groups
* Asian, Scottish Asian or British Asian
* African, Scottish African or British African
* Caribbean or Black
* Other ethnic group

**Accommodation type**

* A whole house or bungalow that is detached, semi-detached or terraced
* A flat, maisonette or apartment that is in a tenement, part of a converted or shared house or in a commercial building
* A mobile or temporary structure

**Tenure**

* Owns with a mortgage or loan
* Owns outright
* Owns with shared equity
* Rents from council or housing association/ Registered social landlord
* Rents from private landlord or letting agency
* Rent from other
* Part rents and part owns
* Lives rent free, Landlord: Council or Housing association/ Registered social landlord
* Lives rent free, Landlord: Private landlord or letting agency
* Lives rent free, Landlord: Private landlord or other

**Activity last week**

* Employed
* Looking/ waiting for work
* Student
* Retired
* Long-term sick or disabled
* Looking after family, other

**Region**

* Dumfries and Galloway, Aberdeenshire, East Dunbartonshire, Highlands, Orkney Islands, Shetland Islands and Na h-Eileanan Siar
* Glasgow
* Dundee City and Aberdeen City
* Argyle and Bute, East Renfrewshire, Moray
* Clackmannanshire, Inverclyde, North Ayrshire, North Lanarkshire, Renfrewshire, West Dunbartonshire and West Lothian
* Angus, East Lothian, Fife, Perth and Kinross, Scottish Borders, South Ayrshire and Stirling
* East Ayrshire, Falkirk, Midlothian and South Lanarkshire
* Edinburgh

# Overcount correction report

The overcount correction was applied to the person level estimates, reducing the national population by 54,297 people from the initial estimates from the model.

# Bias correction report

Corrections for dependence bias were applied in East Renfrewshire, Na h-Eileanan Siar, Inverclyde, Moray, Orkney Islands, South Lanarkshire, Argyll and Bute, Renfrewshire, West Dunbartonshire, West Lothian, Angus, East Dunbartonshire and Glasgow City.

These corrections increased the national household estimates by 12,935 households, and increased the person estimates by 19,439 people.

# Conclusion

This paper has discussed the changes to coverage estimation methodology implemented due to the lower than expected response rate, as well as outlining existing methodology and reporting on the outcome of this process.

The two main changes were a switch to a more flexible and efficient modelling technique based on logistic regression, and the enhancement of our coverage survey with administrative data. We have detailed the reasons that these decisions were made, based on the accuracy and precision of the outputs.

We also outlined how we selected our undercoverage models using a systematic, multistage, validation approach. Finally, we have reported the decisions made in data construction and the final models and data collapsing used.

All our methods, as well as their implementation, have been through extensive peer review, including approval from our international steering group to whom we are very grateful to for their help and advice.

# Glossary

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Response Rate | The response rate is the total number of usual residents whose details were completed on a returned questionnaire, divided by the estimate of the total number of usual residents. |
| Return Rate | Return rates are the number of household questionnaires returned as a proportion of the total active household questionnaires that were in circulation (active refers to all households where the address hadn't been deactivated by the field staff during field operations). Return rates were used during the census field operation to target field staff resources to the lowest responding areas. |
| Census Coverage Survey (CCS) | A survey that takes place after the census to help create more accurate population estimates. |
| Logistic Regression Estimation Model (LREM) | A statistical model that predicts likelihood of responding to census depending on various predictors (e.g. Age, sex, ethnicity). Based on the Census Coverage Survey |
| Office for National Statistics (ONS) | The Office for National Statistics is the executive office of the UK Statistics Authority, a non-ministerial department which reports directly to the UK Parliament. It runs the England and Wales Census. |
| Confidence Intervals (CI) | A confidence interval is a range of values that describes the uncertainty surrounding an estimate. If the estimate was constructed repeatedly (using different CCS samples), the CI describes the range of values that 95% of the estimates would fall in. |
| International Steering Group | A group of census experts from both other producing organisations and academia set up to provide guidance and best practice on Scotland’s Census 2022 |
| Dependence Bias | Dual System Estimation requires two independent sources to produce estimates. When there is some dependence between the two, the estimates produced have bias due to this dependence. |
| Dual System Estimation (DSE) | The statistical process used to estimate the number of people who are missed in the census count. |
| Estimation Area (EA) | A non-contiguous grouping of Local Authorities for use in Estimation. |
| Hard-To-Count Index (HtC) | 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 (LA) | Local Authorities are the 32 council areas in Scotland. |
| Overcoverage/ Overcount | Where individuals are represented within a dataset more than once. |
| Quality Assurance (QA) | Agreed systems that check and validate the work that we do so that our end product is robust and received well by the end users. |
| Ratio Estimator | Part of the Estimation process, which calculates the ratio between the estimates and census counts within CCS sample areas, and applies that ratio to the non-sample census counts to get estimates for the overall population. |
| Undercoverage | Where individuals are not included in the data, and the coverage is less than 100%. |

1. National Records of Scotland. (2023). Statistical Quality Assurance Report, available from <https://www.scotlandscensus.gov.uk/media/awpj2jsa/scotlands-census-2022-quality-assurance-report-for-first-outputs.pdf> [↑](#footnote-ref-1)
2. National Records of Scotland. (2023). Methodology Enhancements to Secure High Quality Census Outputs and Population Estimates. Available from: <https://www.scotlandscensus.gov.uk/documents/methodology-enhancements-to-secure-high-quality-census-outputs-and-population-estimates/> [↑](#footnote-ref-2)
3. For more information see glossary in section 12 [↑](#footnote-ref-3)
4. Račinskij, V. (2018). Coverage Estimation Strategy for the 2021 Census of England & Wales. Available from <https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP105-Coverage-Estimation-Strategy-for-the-2021-Census-of-England-and-Wales.docx> [Accessed 23 May, 2022] [↑](#footnote-ref-4)
5. Our actual error is likely to be higher than this as this assumes a CCS response rate of 70% [↑](#footnote-ref-5)