# Population stock estimates using linked administrative data and a coverage survey – a case study for 2021 and future directions

Eleanor Law, Amy Large, Ceejay Hammond, Mark Linton

September 2023

## Contents

## Key Messages of Paper

Purpose

- This paper summarises our work to estimate the coverage of the Statistical Population Dataset (SPD) version 4 in 2021, using the Census Coverage Survey (CCS). A coverage adjustment is required for the Dynamic Population Model (DPM). We also describe alternative options for future development and set out which we intend to pursue.

Recommendation

- This coverage adjustment is not currently suitable for the DPM to use as there are significant limitations of the estimation of overcoverage and difficulties in reducing the problem to a population excluding large communal establishments that can be estimated effectively.
- A population coverage survey should be designed and implemented to complement the administrative data to enable the ambitious Option 2b.
- In parallel, research should investigate the possible combinations of administrative data and methods for the administrative-only Option 3.
- For both of these options, effective removal of overcoverage cases is required, which may be achieved by modelling the probability of inclusion in the usual resident population using all available data. Audit surveys should be designed and implemented to monitor the effectiveness of such a method, or for rules-based alternatives.

Key Asks of MARP

- We would like the panel to provide feedback on the following:
    - The methods applied within the 2021 case study and the conclusions we have drawn from this. Does the panel agree with our conclusion that the current methods and data available do not produce an effective coverage adjustment?
    - The proposed direction of our work and the importance of producing 2026 population stock estimates independent of the census.

## Executive Summary

- Our previous MARP paper (Law, et al., 2022) outlined high level options for population estimation using Statistical Population Datasets (SPDs), and in this paper we present a basic implementation of Option 1.
- We apply overcoverage estimation and undercoverage estimation to the 2021 SPD v4 by using the 2021 Census Coverage Survey (CCS) as a complementary source.
- These adjusted population stock estimates aim to provide an unbiased coverage adjustment of the SPD for the Dynamic Population Model (DPM).
- We use similar methods to those used in Census estimation, including dual system estimation (DSE).
- Undercoverage is common to both Census estimation and coverage estimation of SPDs, but overcoverage is a much more significant problem in SPD estimation.
- Our methods produce coverage-adjusted SPD estimates for 2021 that are larger than mid-year estimates (MYEs) by 5.9% overall at the national level for England and Wales.
- These estimates are unlikely to be used for the DPM in their current form, as there are major limitations to the methods.
- Several assumptions of DSE are violated when applying the methods to coverage estimation of SPDs.
- We describe the approach we intend to apply for variance estimation, which will provide important quality information for the estimates.
- We believe these methods may not be suitable for SPD coverage estimation using the data currently available.
- Similar methods may be applied in the coming years if better data can be collected and if better methods for overcoverage estimation or trimming can be developed.
- Two options will be pursued in the next phase of our research: one supported by a coverage survey to estimate undercoverage (Option 2b) and one using only administrative data, which is monitored using smaller audit surveys.

## Introduction

In this paper, we describe a specific case study that was used to understand the potential of applying traditional census methods to administrative data in order to estimate coverage and therefore produce population stock estimates. A brief overview of this case study has already been published as part of the June 2023 DPM research publication. We then describe how this case study and other countries' approaches have informed our plans to design and implement a solution for population stock estimation in the future. While the Dynamic Population Model (DPM) can use a Census-based coverage adjustment for the SPD totals in the short term, it has been agreed that an alternative solution for updated stock estimates is required from 2026 onwards. We are working towards an ambitious administrative-based estimation system, possibly supported by a coverage survey to measure undercoverage. We outline our intended plan of research and development for the methods, survey, and data.

The case study aimed to produce a set of coverage weights for the 2021 SPD v4 by sex, age and Local Authority (LA) using the most suitable data currently available, excluding 2021 Census responses. In the future, bespoke survey data collection will likely be in place, which would enable better estimates to be produced. We have previously outlined high level options for population estimation using SPDs (Law, et al., 2022), which we now expand upon:

- Option 1 – estimation of both overcoverage and undercoverage using surveys
- Option 2 –estimating either overcoverage (2a) or undercoverage (2b) using one survey
- Option 3 – a system for estimation that is independent of surveys but may use them for periodic auditing

Here, we have implemented a rudimentary version of Option 1, making use of the high-quality linkage that is available between the SPD and a subset of the Census Coverage Survey (CCS), referred to as CCS2. The results give an indication of the accuracy of population estimates that may be obtained using these methods and without a full Census. We also include a description of the variance estimation method that will be used to estimate the precision of the coverage-adjusted SPD population totals (Annex 5), although this work is not yet complete.

## Background

### Requirements of the Dynamic Population Model (DPM)

For 2016 onwards, the DPM currently takes as population stock estimates the adjusted SPD aggregate totals broken down by LA, single year of age (syoa) and sex. These are reconciled with other stock estimates (if available) and flows estimates via the Bayesian demographic accounting model. An estimate of uncertainty of the SPD totals is also provided to the DPM, using a method based on comparisons between 2011 Census and SPD v3 (previously known as ABPE v3) (ONS, 2020).

Currently, an SPD "coverage adjustment" is used, provided as input information to the DPM. The adjustment is in the form of coverage ratios for each domain (LA by

syoa by sex) that are smoothed using a Generalised Additive Model (GAM) before being supplied to the DPM. Coverage ratios were calculated based on comparisons between earlier versions of 2011 SPDs (v3) to 2011 Census-based mid-year estimates (MYEs). This exercise has been repeated for 2021 Census-based MYEs and 2021 SPD v4. This provides adjusted SPD totals for 2021, assuming that the MYE totals are correct. The adjusted totals are still assumed to have error, modelled as a normal distribution in the data model specified for the stock inputs. From 2021 onwards, the coverage ratios are currently assumed to be constant, which does not reflect reality, in the absence of any other relevant data. Therefore, it is important that the coverage adjustment can be updated at a regular frequency. We are working towards designing and implementing an effective coverage adjustment for the 2026 population stock estimates to be provided to the DPM. Currently Options 2b and 3 are preferred, and we will explain the reasoning behind this in the "Future Work" section.

### Census estimation methods

In some ways, producing estimates of SPD coverage is analogous to producing population estimates using a traditional Census and CCS. There are extra challenges associated with the use of administrative data and the difficulty arising from how it is collected, which are described in more detail in our previous MARP paper (Law, et al., 2022).

The 2011 and 2021 Censuses used slightly different methods to create population estimates, and the methods we have applied here are most similar to the 2011 approach, where stratified dual system estimation (DSE) was used at the level of age by sex by postcode cluster (or Output Area) with a ratio estimator to construct estimates at higher levels of geography (Račinskij, 2018). For overcoverage estimation, overcount propensities were estimated for five-year age-sex groups by region, using the linked Census to CCS data and assuming the CCS determined the correct location of Census individuals.

For Census 2021, DSE was carried out using mixed effects and fixed effects logistic regression models, enabling pooling of data across geographical areas and the use of relevant covariates, e.g. tenure and household size. Overcoverage estimation was carried out at the national level, again using logistic regression and the inclusion of relevant statistically significant covariates (ONS, 2022). Currently, when working with SPDs and the CCS, we do not have access to the same kind of high-quality covariates on the SPD that were used in Census 2021. For the case study presented in this paper, stratified DSE (similar to Census 2011) was chosen, which is explained more fully in the Methods section.

### Application to SPDs

DSE has previously been applied to 2011 SPD v2, using deterministic matching to Census 2011 data and simulating a 1% population coverage survey from the Census data (ONS, n.d.), however, no real survey data were used. Using the original SPD v2, and also applying an overcoverage adjustment similar to that used in Census 2021, a relative bias of 7.7% was achieved. When a model using a combination of variables was used to derive a score and remove cases likely to be overcoverage,

the bias was reduced to 3.7%, but false negative links were thought to be another main reason for over-estimation.

As demonstrated in this previous work, overcoverage is a much more substantial problem for population datasets constructed from administrative data than it is for traditional census returns. A census return clearly establishes the usual residence of an individual, but some interactions, for example with health data, may take place even if England and Wales are not an individual's place of usual residence. Estimation of overcoverage, in the way that it has been done for census, depends upon flagging individuals as overcoverage. Cases are picked up as overcoverage on the CCS, or by census-census linkage in the case of duplicates. Erroneous inclusion in the SPD includes people who were not usually resident in the UK on the SPD reference date. This overcoverage cannot be estimated or modelled without data on those cases, which is difficult to obtain.

Currently, the best available strategy to reduce overcoverage is to use "trimmed" DSE, which has been tested by CSO Ireland as a method to compile estimates using admin data only (CSO, 2021). Records are scored on how likely they are to be erroneous. The records can be scored on a number of different parameters. In the work using Irish admin data, the authors use income to score records. This assumes that records with very small incomes are more likely to be erroneously included.

There are some risks associated with trimming. DSE assumes homogeneous capture within strata, i.e. equal capture probability, in one of the lists. If this assumption does not hold, the resulting bias in estimates depends on whether heterogeneity is aligned to any heterogeneity in the other list. However, of those individuals who are in the usually resident population, trimming is more likely to remove some types of people than others, and the same people may be underrepresented in any second list used, whether that is a survey or another administrative list.

In this paper, we describe the data and methods that we considered to be the best practically available to produce a realistic set of coverage estimates for the SPD in the timeframe that was required for the June 2023 DPM publications. The limitations, and quality achieved with this data are discussed, together with possible improvements. We also describe our plan to build on this work to produce a census-independent coverage estimation process that can be used to provide population stock estimates from 2026 onwards to feed into the DPM.

## Data

### Statistical Population Dataset (SPD)

We used the 2021 SPD v4, which uses 30 June as the reference date and was constructed using the Demographic Index v2.1, which provides the linkage between the constituent data sources of the SPD. SPD v4 uses three new administrative datasets in addition to those on SPD v3, but a similar methodology of inclusion rules based on activity within the last year (ONS, 2023). Therefore, v4 reduces undercoverage compared to v3 but still contains overcoverage cases.

To reduce overcoverage, we applied a process of ‘trimming’ (removing) records in the SPD for which we have weaker evidence that they are in the usually resident population. Some of the variables from the construction of the SPD may give an indication of the confidence in that record being correctly included in the population. Those that capture the date of the last interaction with an administrative list may be used to add stricter criteria for inclusion in the dataset. This process of trimming less certain records used the variables listed in Annex 1 and is described in Annex 3.

## Census Coverage Survey 2021 (CCS)

A Census Coverage Survey (CCS) was run in 2021 eight weeks after Census day (21st March) as a second capture to be used in the DSE (ONS, 2022). The clustered design of the CCS was optimised to produce the most precise estimates possible from the 2021 Census, by oversampling Output Areas (OAs) that had a lower rate of census returns in 2011 (i.e. had a higher “hard to count” index) (Burke & Račinskij, 2020). The primary sampling unit (PSU) was the 2011 OA, sampled from strata of LA by hard-to-count, where the number of OAs sampled was optimally allocated between strata. The secondary sampling units were postcodes, sampled from each OA at a set sampling fraction of 25%. Postcodes were excluded from the sampling frame where they had no residential addresses or if they were entirely occupied by residences classed as Large Communal Establishments (LCEs) (capacity of more than 50 bed spaces). The differing probabilities of selection of a postcode into the CCS sample were used to derive sampling weights for the CCS that were used in the modelling for the 2021 Census estimation.

The Census and CCS were linked using automatic probabilistic and additional clerical linkage to satisfy the high-quality requirements of Census estimation. An additional linkage exercise was carried out to link the Census/CCS dataset to the Demographic Index (DI), in order to better understand the overcoverage and undercoverage of the DI and of SPDs, which are derived from it (ONS, 2023). With the available clerical resource, it was only possible to link a subset of the Census and CCS data to the DI. Therefore, a subsample of the CCS was selected, stratified by OA to approximately maintain the proportions by LA and hard-to-count in the original CCS sample. This subsample is referred to as CCS2.

## Large communal establishments (LCEs)

As the CCS does not collect data from LCEs, we used data from Census 2021 to label cases on the SPD that were in LCEs for removal from the SPD, along with any linked record on CCS, for the few cases where this applied. We used two lookup datasets from the 2021 Census LCE estimation to flag records that were placed at LCE addresses on the SPD (See Annex 2 for details). 41% of records in the SPD do not have a unique property reference number (UPRN) because the administrative sources for that record do not provide UPRN. We were unable to remove any of these records, therefore adding LCE estimates back on at the end of estimation will result in over-estimation.

We used final estimated LCE populations from Census to add to our LCE-excluding estimates so that they could be compared to MYEs.

2021 Mid-year estimates (MYEs)

We used published Census-based 2021 MYEs for England and Wales (ONS, 2022) as a reference to determine the error in our final coverage-adjusted SPD estimates. These estimates also have associated uncertainty (ONS, 2012-2016).

## Methods

First, we defined the population to be estimated using these data sources, to establish the coverage of the SPD for that population of interest. We aimed to estimate the coverage of the SPD for private households and small communal establishments (SCEs, 7-49 bed spaces). Because our second list, the CCS, is not designed to cover LCEs, we excluded LCEs from the population to be estimated. The CCS did include some records from SCEs, but we would not expect them to have the same probability of inclusion in CCS as residents of private households. Ideally, we would estimate SCEs separately because of this, but to simplify the estimation and due to concerns about small sample counts, in this case we grouped records together from private households and SCEs. In our analysis, we describe the coverage adjustment that is applied by age groups and sex and LA in this population of interest. We also made comparisons to mid-year estimates, which do include LCEs and therefore we added LCE estimates to our population of interest to make this comparison.

Our aim was to create unbiased estimates of this population using an undercoverage adjustment and an overcoverage adjustment, as the SPD is known to have both.

Steps were carried out in the following order:

1. Remove cases from SPD that are placed at addresses labelled as LCEs
2. Carry out estimation steps (Figure 1) using LA by sex by five-year age band as strata. These steps are described in detail in Annex 3.
3. Add LCE totals to estimated totals for private households and SCEs for comparison to MYEs

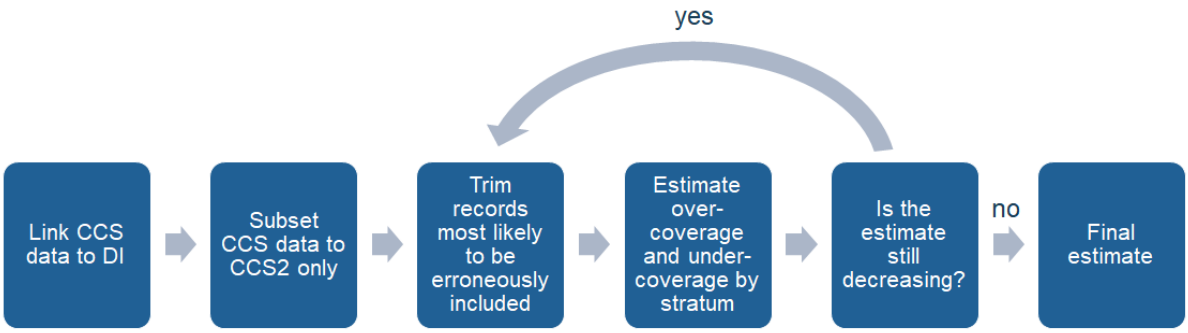


*Figure 1: Estimation steps*

Additional methods that form part of our ongoing work are described in Annexes 4 and 5.

## Results

We produced “coverage-adjusted SPD estimates” of the June 2021 population of England and Wales using trimming, stratified DSE and weighted overcoverage estimation as described in the Methods section and Annex 3.

### Comparison to mid-year estimates (MYEs)

The coverage-adjusted SPD estimates were compared with the Census 2021-based MYEs to measure the coverage error. We treat the Census 2021-based MYEs as correct, given that the time elapsed since 2021 Census is minimal and therefore uncertainty due to coverage drift is small. The total coverage adjusted SPD estimate for England and Wales was greater than the MYE by 3.98%. This is greater than the difference between the unadjusted SPD count and the MYE. Without applying trimming, the over-estimation is greater at 5.36%. We only present estimates using trimming in the following results. We do not include measures of uncertainty for our coverage adjusted SPD estimates here as our work to estimate sampling variance is still in progress (see Annex 5 for our intended method).

*Table 1: National (England and Wales) population sizes by method*

| Method | National (England and Wales) June 2021 population total | Difference relative to MYE (%) |
|---|---|---|
| **Census 2021-based MYE** | 59,641,829 | 0 |
| **Unadjusted SPD count** | 58,949,900 | -0.77 |
| **Coverage adjusted SPD estimate** | 62,017,780 | 3.98 |

Coverage error differed across age groups, as shown in Figure 2. The oldest and youngest age groups had the smallest coverage error, and males had higher coverage error than females except for the 10 to 14 and 15 to 19 years categories.

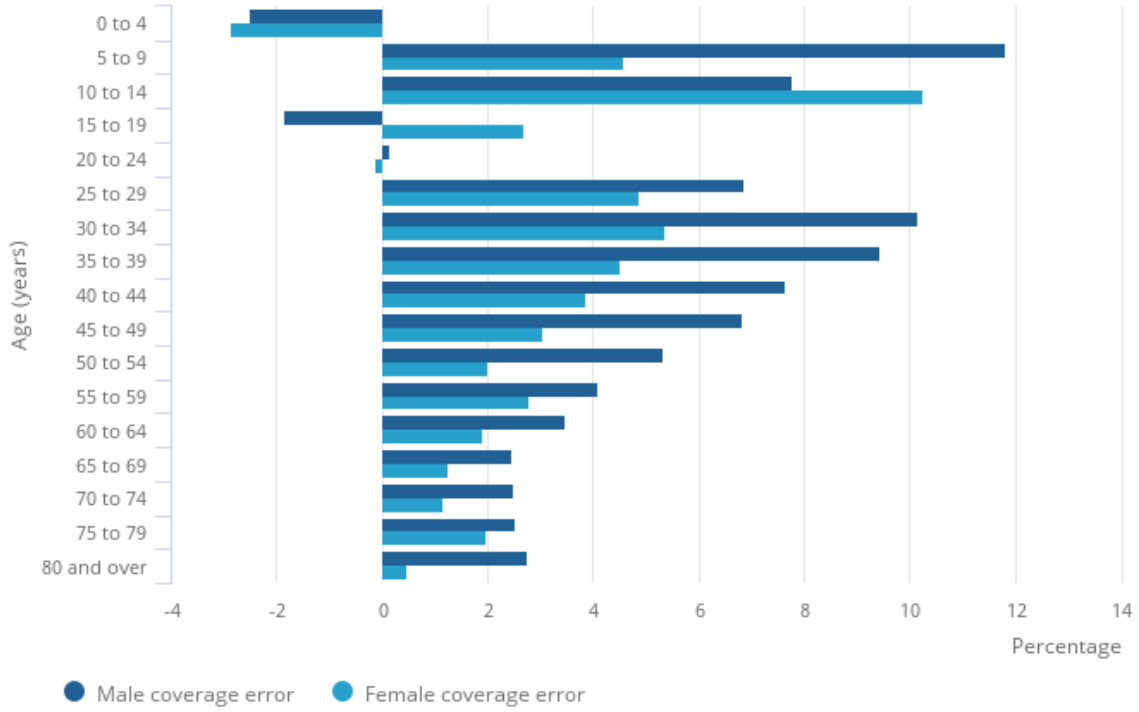


*Figure 2: Coverage error of coverage-adjusted SPD estimates, using 2021 Census-based MYEs as the reference population size assumed to be true, by five-year age group and sex*

Figure 3 shows the coverage error by local authority as a percentage of the local authority Census 2021-based MYE population size. Most local authority estimates had a positive coverage error between 0 and 10%, that is, the coverage adjusted SPD estimates were greater than the MYE. A small number of local authority estimates had negative coverage error.

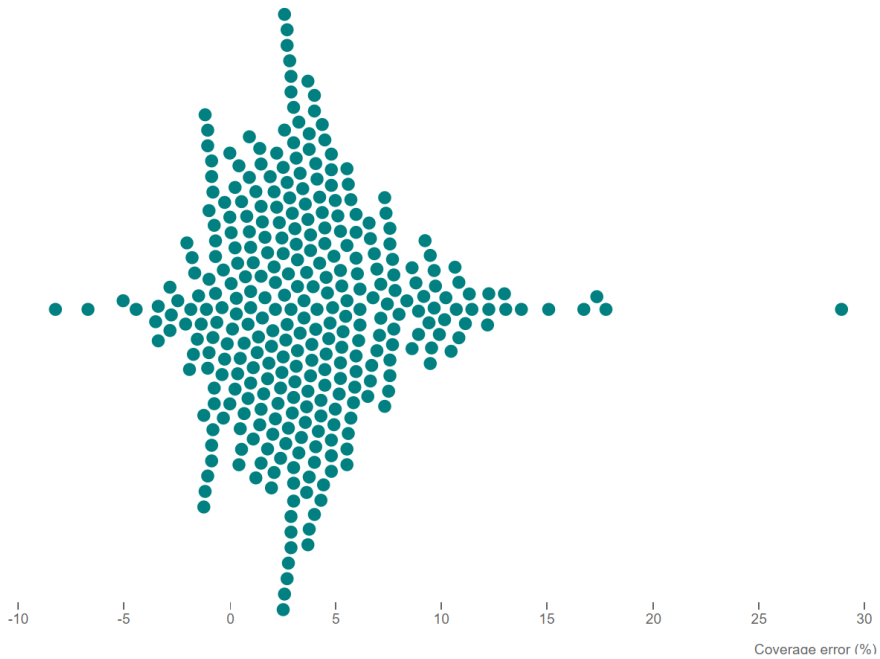


*Figure 3: Local authority-level coverage error of coverage-adjusted SPD estimates, using 2021 Census-based MYEs as the reference population size assumed to be true. The local authorities are ordered on the x-axis by coverage error.*

Figure 4 and Figure 5 show the geospatial distribution of coverage error by local authority. Urban areas, especially London, often had greater positive coverage error. City of London was an extreme outlier (coverage error 29%) and is not shown on this scale so that the differences between other local authorities are more visible.

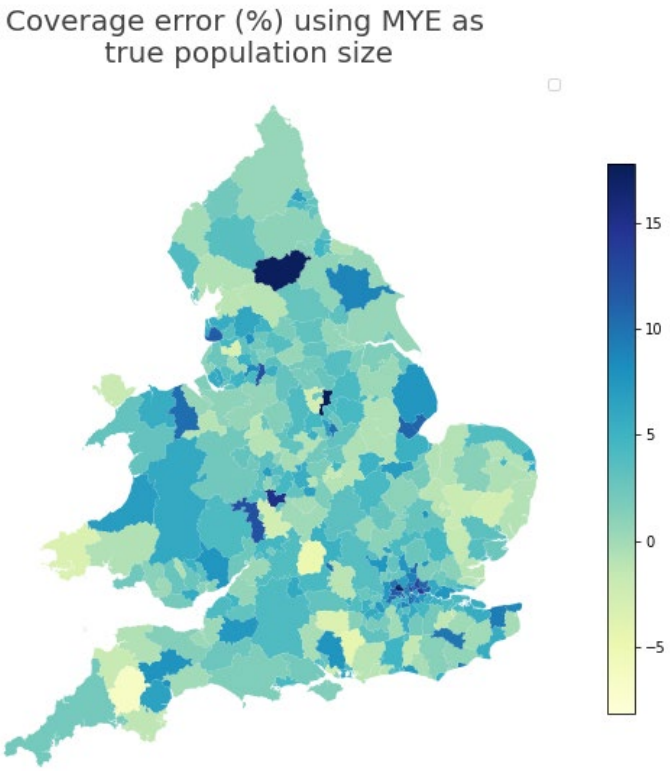


*Figure 4: Local authority-level coverage error of coverage adjusted SPD estimates, using 2021 Census-based MYEs as the reference population size assumed to be true.*

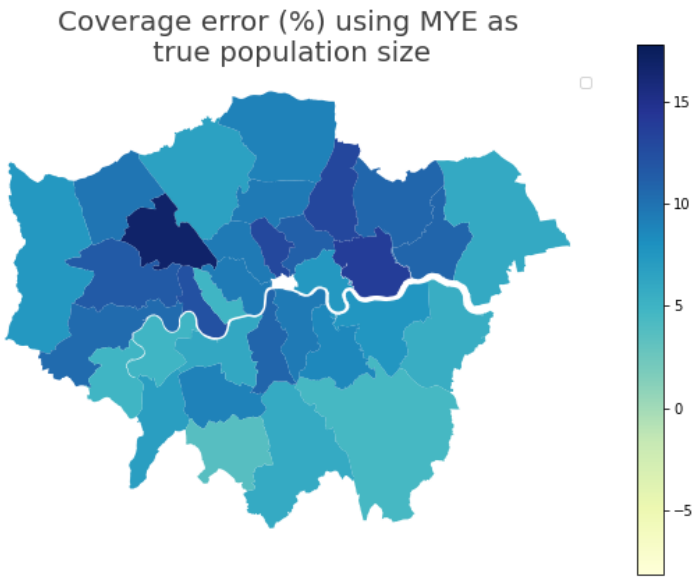


*Figure 5: Greater London from Figure 4, shown with a larger scale. City of London (central white area) is not shown so that the differences between other local authorities are more easily observed.*

## Discussion

Our results show a large difference between the coverage-adjusted SPD estimates and the Census-based 2021 MYEs, which would be expected, given the considerable limitations of the method as it currently stands. We should also bear in mind that MYEs have some error (ONS, 2012-2016), but this error is much smaller than the differences observed.

A larger adjustment was applied for undercount than for overcount, for almost all strata of estimation. We believe that undercoverage estimation was able to adjust well for individuals omitted from or misplaced (located in the wrong LA) in the SPD. However, overcoverage estimation was dominated by misplacement, with other important types of overcoverage not accounted for. A small number of non-usual residents responded to the CCS, and therefore we used those cases to estimate their contribution to overcoverage, but it is not reasonable to assume that non-usual residents responded to CCS at the same rate as usual residents. People who emigrated since interacting with a service in the previous year, or people who briefly visited the UK, interacted with a service and then left, had no probability of responding to CCS. These people were therefore unaccounted for in our estimation methods as we had no data to label them as overcoverage.

A subset of these overcoverage individuals were removed by the trimming method, but the overcoverage that remained still led to bias in the final estimates. Trimming reduced coverage-adjusted SPD estimates by a few percent compared to those without trimming. Further work is required to improve the method to the point where over-coverage becomes “negligible”, or so small across all areas that a set adjustment is acceptable. Other countries have used modelling methods or rules-based “signs of life” methods.

### Addressing overcoverage

Overcoverage in the SPD is the most significant challenge in providing unbiased population stock estimates to the DPM. There are two proposed solutions to address this (ONS, 2023).

The first solution is to collect accurate data on individuals who are leaving or have left the usually resident population. The greatest challenge is differentiating non-responders from people no longer resident. To do this to the highest quality would require additional fieldwork. There are two main ways to use administrative data while carrying out such fieldwork:

- **Dependent sampling** is where the sample is drawn using information from an administrative dataset, such as in this case the SPD
- **Dependent interviewing** is where interviewers share information from administrative datasets with respondents to either verify or correct it. However this approach was not approved last time it was discussed by the national statistician's data ethics advisory committee (NSDEC).The SPD also does not hold contact details to enable follow up apart from addresses. The SPD also does not hold contact details to enable follow up apart from addresses.

The second solution to the overcoverage problem is to use stricter rules or a model, taking advantage of more data sources, to produce an SPD or administrative lists with less overcoverage. The aim would be to remove individuals from the SPD until negligible overcoverage remains. This inevitably increases undercoverage, but the aim is not to produce an SPD of a similar size to the true population. When applying trimming in the 2021 case study previously, we did not use any data aside from the core SPD variables. Other sources and variables could be considered. The linked 2021 Census and administrative data may enable us to train a model or develop rules that remove more overcoverage, even if this increases undercoverage. This may work well in the short-term but would require review and audit. If no other sources are available, we would require data as described in the first solution.

### Identifying the population of interest

In addition to problems with overcoverage, it is also currently very difficult to restrict the scope of estimation to the population we defined (private households and SCEs only). The census data from 2021 that provides estimates of occupancy is of high quality and completeness. However, to use it to remove cases from the SPD, we relied on UPRN, which is not available for the 41% of SPD records that do not have an English Schools Census or Patient Demographic Service record. In Census 2021, 1.7% of the population were estimated to live in communal establishments, and the majority of those live in LCEs. If the 41% of SPD records that lack UPRN have the same proportion in LCEs as those with UPRN, estimates would be inflated on average by no more than 1%. In reality, LCEs are not evenly distributed by geography and their residents are more likely to be student age or elderly. Therefore, in some LAs there will be greater overestimation of these groups. LCE residents are possibly more likely to be using health services and therefore have a UPRN on the SPD, which would slightly reduce the overestimation. We also had to assume that there were no responses to CCS from LCEs as this incorporated into the design, even though a small number (<1000) were removed as they did link to SPD records that were labelled to be in LCEs.

Future SPDs will be built using Frameworks data from HMRC instead of CIS data from DWP. Frameworks should provide UPRN for the vast majority of SPD cases that do not have PDS or ESC, and therefore this should make it much easier to carry out estimation of coverage for UPRN-labelled subpopulations of the SPD, such as excluding LCEs. However, partitioning the population into groups based on the type of residence or special populations will be difficult, and individuals who move between populations that are estimated in different ways will bring new challenges of overcoverage and undercoverage.

### Problematic assumptions

In addition to these fundamental problems with the coverage estimation we have carried out, there are some assumptions required when using DSE that do not hold in the context of the SPD and CCS2 data that we used. Similar assumptions were made when carrying out estimation for Census, but when using SPDs instead, we have less control to reduce the impact of violating them.

It is assumed that one of the lists has equal capture probability for all members of the population of interest. We require this to be true for the CCS, as the SPD will certainly cover some types of individual better than others e.g. 19 year olds working or attending university compared to their peers volunteering during a gap year. Many factors will also affect a person's probability of responding to CCS, therefore estimates may be biased in either direction depending on how this is manifested in the data. When we apply trimming, we also further distort the representativeness of the SPD, which exacerbates the impact of the CCS not having equal capture probability for all people.

It is assumed that the population is closed. We know this assumption is violated as we know that there is immigration, emigration, and movement in and out of LCEs (which were excluded from the population) between when the SPD record data and CCS responses were collected. For Census, the date of the CCS is set as close as possible to Census day whilst still maximising their independence, and people are specifically asked where they were on Census day. For our estimation, we must use the address on Census day, which may be up to nine months after or three months before the administrative record that qualified the person to be included in the SPD (the 2021 June SPD uses records from the previous July onwards). Movements between these dates will lead to overestimation, as emigration and immigration will appear as non-response to CCS or failure to capture someone on the SPD.

It is assumed that populations are homogeneous within strata. We hope to improve the undercoverage adjustment using smaller strata to improve the validity of this assumption. However, for overcoverage estimation, we require larger groupings in order to have sufficient counts. In this work, we used LA supergroups, which should be an improvement on national estimation, but each supergroup still contains very diverse LAs. The effect of carrying out overcoverage estimation at a level higher than LA is that even if on average the estimation is reasonable, at the LA level there will be some that are over-estimated and others that are under-estimated because the variation has been averaged over a supergroup. Using a logistic regression model would be preferable, but it would be necessary to identify suitable variables for modelling from the SPD or that can be joined to the SPD, like those used in fractional counting (ONS, 2023). All types of overcoverage are estimated together due to sample counts, but they would be better modelled separately, as they are likely to be associated with different kinds of people.

## Estimation Options

Since the June publication, while revisiting the methods for estimation of overcoverage from 2011 Census (particularly with regards to misplacement), we have also noted that the cases used to estimate misplacement overcoverage do not appear to be representative of all such cases when using the sampling weights in the way we currently do. We will explore calibrated DSE as this framework should account for these cases properly, and give us confidence that only national level overcoverage needs to be removed by our trimming method (Zhang, 2023).

We now extend information on the options briefly described in the "Future options for Coverage Estimation" section of the June DPM Research publication (ONS, 2023).

Some approaches use surveys to estimate both over- and/or undercoverage error. Some require periodic auditing surveys to monitor over- or undercoverage error. These approaches are what we currently consider to be the most viable options for producing coverage-adjusted population estimates.

Option 1: Two surveys are used for estimation and the population dataset (SPD) is used in its current from.

- Survey 1: An area-or-address-based sample survey is used to estimate undercoverage error.
- Survey 2: A list-based dependent sample survey with dependent interviewing is used to estimate overcoverage error.

An example of this approach being implemented is for the Italian Population and Housing Permanent Census, which makes use of an area-based and list-based survey to estimate coverage error in the Population Base Register (Bernardini, et al., 2022).

Option 2: One survey is used for estimation (two different sub-options).

*Option 2a:*

- A population dataset where undercoverage error is considered negligible. An example of population dataset could be the Demographic Index (DI).
- An inclusion model is used to estimate overcoverage error of the lists and dependent sample and interviewing is required to update the models.
- A periodic area or address-based sample may be required for auditing undercoverage error.

Israel currently uses a system that is a variant of this option, which made use of combining data from the population register with data collected in the field from a list-based survey to estimate overcoverage error (Pfeffermann, et al., 2019).

Option 2b:

- A population dataset where overcoverage error is considered negligible by using strict inclusion rules or model-based trimming. In this case the population dataset could be the SPD in its current form.
- An area- or address-based sample survey is used to estimate undercoverage error.
- A periodic list-based dependent sample survey and dependent interviewing may be required to audit inclusion rude or model-based trimming.

Option 3: Estimation system and Audit Survey

No survey is used for estimation. Surveys are used only for auditing. The administrative sources used may be combined as in the current SPD or kept separate.

Two or more separate administrative lists recording interactions with different services (or a combination of services) are used in an estimation system. One of these may be constructed in a similar way to the current SPD. In the case of more

than two lists, inter-dependence may be modelled. This option is dependent on undercoverage being minimised and effectively estimated using multiple lists. Overcoverage is reduced to a negligible level by using strict inclusion rules or model-based trimming scores. Ongoing surveys would be required to collect data to audit undercoverage, accuracy of inclusion rules or model-based trimming and correct placement.

An example of this is the Population Estimates Compiled from Administrative Data Only (PECADO) approach described by Dunne and Zhang (2023). This approach made use of trimming the population register of erroneous records and then linking this register to Driving Licence Data (DLD), where the level of overcoverage error is deemed to be negligible. The Trimmed Dual System Estimator (TDSE) is then applied.

## Conclusion

The net overcoverage of the coverage-adjusted SPD estimates is 3.98% for England and Wales, when compared to Census-based 2021 MYEs as a reference, and some LAs are overestimated by over 15%. This magnitude of bias means that the estimates in their current form are unlikely to be useable, but it does provide a useful indication of the quality that may be achieved with similar types of data and methods in the future.

We believe that although small improvements can be made to this type of method, it is not a long-term solution to the requirement for an unbiased coverage adjustment method for the SPDs that will feed into the DPM. However, extensions to the current methodology, such as improving trimming, implementing one of the discussed options for a second source to measure coverage and the consideration of the calibrated DSE could address these limitations.

---

Does the panel agree with our conclusion that the current methods and data available do not produce an effective coverage adjustment?

---

## Future Steps

In the short term, we aim to implement the method described using the DSE with stratification at a lower level of geography, Ratio and Local Synthetic estimators described in Annex 4.

Once we have implemented the updated estimation method, we will rerun the implemented variance estimation methods described in Annex 5 to provide estimates of precision.

For similar work in the future, it should be possible to better identify population groups within the population dataset. Currently we are collaborating with colleagues to develop methods to estimate special populations, which will enable us to partition the population into different population groups for estimation. Identifying CEs will be

an important part of this and may be facilitated by better geographic precision, provided that Frameworks data replaces DWP data as a source of income information.

In our longer term work we will focus on the development and implementation of Options 2b and 3. These are the preferred options because firstly, we have operational experience of delivering a high quality survey specifically designed to measure and adjust for undercoverage, and the methodology needed to apply these approaches is well defined and understood from the experience gained from previous census work. Secondly, an administrative only option aligns with the organisational objectives of being radical and progressive, and also potentially offers better value for money in the long term. This will be the case if the administrative data can be shown to be stable with sustainable delivery and governance, and that methods can be developed and quality assured that work with the available data.

A rolling coverage survey would enable Option 2b to be carried out annually, but the design of this survey, and whether it is independent of the Transformed Labour Force Survey (TLFS) is to be confirmed. Alternatively, a higher quality CCS-like coverage survey may be run every five years, for example.

The development of these methods also requires stricter inclusion rules and model-based trimming to remove erroneous records from the population datasets. These options depend less on the type of reliable overcoverage data that may be difficult to collect using dependent sampling and interviewing, which also have ethical concerns that will need to be addressed. The use of additional administrative data sources, e.g. border and visa data, will also be explored to complement the use of rule- or model-based trimming. Trimming reduces the size of available data for estimation, therefore it will be important to maximise the accuracy with which potential overcoverage is flagged. This work will build on the inclusion modelling previously developed as “Stage 1” of the Fractional Counting work.

We are currently intending to use the 2021 Census based coverage ratios for input into the DPM until the production of 2026 estimates. This will allow us an opportunity to develop our tactical coverage assessment process (via survey, Option 2b) whilst also quality assessing and testing administrative data sources which are vital in the successful implementation of Option 3. If Option 3 cannot be developed to a suitable level of quality within this time frame, the Option 2b results will serve as a suitable placeholder until an administrative data only approach can be constructed and quality assured.

---

Does the panel believe that the approach of developing both an admin only and a fallback using survey is a sensible one?

Does the panel feel the two options (2b and 3) we are focussing on are the right ones?

Does the panel have any guidance on implementing overcoverage assessment given ethical concerns about dependent sampling and / or interviewing?

_____

## References


Abbott, O., 2009. 2011 UK Census Coverage assessment and adjustment methodology. *Population Trends,* Volume 137.

Bernardini, A. et al., 2022. Evolution of the person census and the estimation of population counts in New Zealand, United Kingdom, Italy and Israel. *Statistical Journal of the IAOS,* pp. 1-17.

Burke, D. & Račinskij, V., 2020. *2021 Census coverage survey: sample allocation strategy.* [Online]
Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP127-CCS-2021-allocation-strategy.docx

Chapman, D. G., 1951. *Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses.* s.l.:University of California Press.

CSO, I., 2021. *Irish Population Estimates from Administrative Data Sources, 2020.* [Online]
Available at: https://www.cso.ie/en/releasesandpublications/fp/fp-ipeads/irishpopulationestimatesfromadministrativedatasources2020/

Dunne, J. & Zhang, L., 2023. A system of population estimates compiled from administrative data only. *Journal of the Royal Statistical Society: Series A (Statistics in Society).*

Law, E. et al., 2022. *SPD Estimation Options.* [Online]
Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2023/02/EAP184-SPD-Estimation-Options-MARP.pdf

ONS, 2012-2016. *Methodology for measuring uncertainty in ONS local authority mid-year population estimates: 2012 to 2016.* [Online]
Available at:
methodologyformeasuringuncertaintyinonslocalauthoritymidyearpopulationestimates2012to2015

ONS, 2020. *Indicative uncertainty intervals for the admin-based population estimates: July 2020.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/indicativeuncertaintyintervalsfortheadminbasedpopulationestimatesjuly2020

ONS, 2022. *Coverage estimation for Census 2021 in England and Wales.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/pop

ulationestimates/methodologies/coverageestimationforcensus2021inenglandandwales

ONS, 2022. *Population estimates for the UK, England, Wales, Scotland and Northern Ireland: mid-2021.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021

ONS, 2023. *Developing Statistical Population Datasets, England and Wales: 2021.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingstatisticalpopulationdatasetsenglandandwales/2021#data-sources-and-quality

ONS, 2023. *Dynamic population model, improvements to data sources and methodology for local authorities, England and Wales: 2021 to 2022.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/dynamicpopulationmodelimprovementstodatasourcesandmethodologyforlocalauthoritiesenglandandwales2021to2022
[Accessed 2023].

ONS, 2023. *Fractional counting: a method to fractionally weight and count integrated administrative data for population statistics.* [Online]
Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2023/02/EAP188-Fractional-Counting.pdf

ONS, 2023. *Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/understandingqualityofthestatisticalpopulationdatasetinenglandandwalesusingthe2021censusdemographicindexlinkage/2023-02-28

ONS, n.d. *Research Outputs: Coverage-adjusted administrative data population estimates for England and Wales, 2011.* [Online]
Available at:
https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/censusanddatacollectiontransformationprogramme/futureofpopulationandsocialstatistics/methodology/researchoutputscoverageadjustedadministrativedatapopulationestimatesforenglandandwales2011

Pfeffermann, D., Ben-Hur, D. & Blum, O., 2019. Planning the next Census for Israel. *Statistics in Transition,* Volume 20(1), pp. 7-19.

Račinskij, V., 2018. *Coverage Estimation Strategy for the 2021 Census of England and Wales.* [Online]
Available at: https://uksa.statisticsauthority.gov.uk/wp-

content/uploads/2020/07/EAP105-Coverage-Estimation-Strategy-for-the-2021-Census-of-England-and-Wales.docx

Wolter, K., 2007. *Introduction to variance estimation.* 53 ed. New York: Springer.

Zhang, L.-C., 2023. *Calibrated trimmed dual system estimation.* s.l.:Internal, available upon request.

## Annex 1 – SPD variables used for trimming

[not available for this publicly available version]

## Annex 2 – Removal of Large Communal Establishments (LCEs) from SPD

The first lookup aimed to update unique property reference numbers (UPRNs) on the SPD for grouped LCEs such as universities. This is because in the Address Frame used for Census, some UPRNs were misclassified to be households and were found to be LCEs in the Census. Therefore, dummy UPRNs were created to combine these misclassified UPRNs, which were then joined onto the SPD. A lookup of LCEs from the 2021 Census (with misclassifications resolved) was then joined onto the SPD using the previous lookup so LCEs could be flagged on the SPD. The steps for this are as follows:

- Attach UPRN values onto the SPD using PDS and ESC, joining on ID numbers
- Update UPRN values in the SPD using grouping datasets. This is done using the first lookup with included original UPRNs (some of which were misclassified) to join onto the SPD and the corresponding Census UPRNs.
- Manually update some missing UPRN values in the Large CE census dataset for completeness
- Join the 2021 Census Large CEUPRNs onto the SPD using UPRN (including dummy UPRNs). From this a Large CE flag can be created for records whose UPRNs are in common with the census lookups or not.
- Add flag to label each column as Large CE or not

## Annex 3 – Estimation methods

We describe these methods not in the order they are used, but instead starting from the most basic method that could be applied to then building on that using various modifications and enhancements that are used in addition.

## Dual System Estimation (DSE)

Stratified DSE was used to estimate undercoverage. Estimates were therefore derived for each sex by five-year age band by LA by Hard to count. For a given LA, the two linked lists of individuals that were used for estimation were

- List A: the list of SPD individuals placed in the LA
- List B: the list of individuals responding to CCS from a CCS2 postcode who stated that they were living in that LA on Census day

If, in the DI to Census/CCS linkage exercise, a record on list A and a record on list B are considered to be the same person, and they are placed in the same LA on both lists, they are considered to be captured on both list A and list B in the LA of interest.

Using the Chapman correction (Chapman, 1951, pp. 60-131) to account for zero and small counts, the dual system estimator $\widetilde{N}_h$ is defined in the following way for stratum *h*:

$$\widetilde{N}_h = \frac{(n_{h_{1+}} + 1)(n_{h_{+1}} + 1)}{n_{h_{11}} + 1} - 1 \qquad (1)$$

where $n_{1+}$ is the total number of individuals captured on List A; $n_{+1}$ is the total number of individuals captured on List B; and $n_{11}$ is the number of individuals captured on both list A and list B.

*Table 2: Table to show the notation used in Equation 1. Observed counts are unshaded; unobserved (estimated) counts are shaded grey*

| | | CCS (List B) | | |
|---|---|---|---|---|
| | | In | Missed | Total |
| SPD (List A) | In | $n_{11}$ | $n_{10}$ | $n_{1+}$ |
| | Missed | $n_{01}$ | $n_{00}$ | $n_{0+}$ |
| | Total | $n_{+1}$ | $n_{+0}$ | $n_{++}$ |

## Overcoverage propensity groups

Overcoverage was estimated as far as possible where it was identified by data collected from CCS and available in the CCS2 subset. We adjusted for it by calculating a simple overcount weight $\hat{g}_{as}$ (and its inverse, $\hat{\gamma}_{as}$), by supergroup *s* and age-sex group *a* as the ratio

$$\hat{g}_{as} = \frac{1}{\hat{\gamma}_{as}} = \frac{n_{as_t}}{n_{as}} \qquad (2)$$

where $n_{as}$ is the total population count for the individuals on the SPD placed in supergroup *s* and age-sex group *a* and linked to a record in CCS2; $n_{as_t}$ is the true population count for the individuals on the SPD placed in supergroup *s* and age-sex group *a* and linked to a record in CCS2 in the same LA. Strata for DSE are nested within supergroups so that for all individuals within a stratum, the same $\hat{\gamma}_{as}$ applies.

To determine which individuals are counted as "true population", three types of overcoverage are excluded (i.e. included $n_{as}$ but not $n_{as_t}$):

1. Misplacement overcoverage is defined as individuals on the SPD placed in the LA being estimated who are on the CCS in a different LA.
2. Duplication overcoverage is defined as two or more individuals on the SPD linked to a single CCS response.
3. Non-usual residents overcoverage is defined as individuals captured in CCS responses as non-usual residents (for example short-term migrants)

The overcount propensity is applied to adjust $n_{h_{1+}}$ in Equation 1 to give the following:

$$\widetilde{N}_h = \frac{\left(\hat{g}_{as} n_{h_{1+}} + 1\right)\left(n_{h_{+1}} + 1\right)}{n_{h_{11}} + 1} - 1 \tag{3}$$

## CCS2 sampling weights

We used sampling weights for overcoverage estimation. Using sampling weights corrected for the higher probability of inclusion of some postcodes in CCS2, which could otherwise introduce bias into the estimation. The different probabilities of inclusion are driven by the different sampling probabilities by LA and hard to count, and since the strata for undercoverage estimation incorporate these, we did not use weights for undercoverage estimation.

The sampling weight $w_{pi}$ for individual $i$ in postcode $p$ was calculated as follows:

$$w_{pi} = \frac{d_p P_a}{p_a} \tag{4}$$

where $d_p$ is the design weight of the postcode that was previously calculated using information about the CCS sample design; $P_a$ is the total number of postcodes in OA *a* in the CCS sample, and $p_a$ is the number of postcodes selected in area *a* in the CCS2 subsample.

The overcount propensity, $\hat{g}_{as}$, was adapted to use the CCS2 sampling weights:

$$\hat{g}_{as} = \frac{1}{\hat{\gamma}_{as}} = \frac{\sum_{i \in M_t} w_{pi}}{\sum_{i \in M} w_{pi}} \tag{5}$$

where *M* is the population of individuals matched to CCS2; $M_t$ is the subset of *M* that in the "true population" (defined in the "Overcoverage propensity groups" section).

## Trimming

We used trimming to remove overcount from the SPD. When trimming, records were removed, starting with those that are most likely to be erroneous. Estimation was then repeated, and this process was iterated, removing a greater number of possibly erroneous records each time, as described in these steps:

1. Score records on a given parameter that may be correlated with overcoverage (e.g. date of interaction or household income).
2. Remove the records with the worst score (oldest date of interaction or lowest income)

3. Run the full estimation process with the trimmed list.
4. Compare new estimates with estimates from untrimmed list. If trimming is removing overcoverage, the estimates from the trimmed list should be smaller than the estimates from the untrimmed list.
5. Repeat the process, removing more and more records.
6. Stop trimming when the estimates converge, that is, they stop decreasing. At this point, true records are as likely to be removed as overcoverage records.

If list A (the SPD) is $n_{1+}$, the naïve DSE (including the Chapman correction) would be

$$\widetilde{N}_h = \frac{\left(n_{h_{1+}} + 1\right)\left(n_{h_{+1}} + 1\right)}{n_{h_{11}} + 1} - 1 \tag{1}$$

If we could remove the overcoverage, and if $r_h$ is the number of records that are overcoverage in stratum $h$, the new estimate, $\dot{N}$, would be

$$\dot{N}_h = \frac{\left((n_{h_{1+}} - r) + 1\right)\left(n_{h_{+1}} + 1\right)}{n_{h_{11}} + 1} - 1. \tag{7}$$

As $r > 0$, $\widetilde{N}_h > \dot{N}_h$, and thus $\widetilde{N}_h$ is an overestimate. Unfortunately, we do not know which records are overcoverage, so we cannot simply remove $r$.

If the number of records removed due to trimming in stratum $h$ is $k_h$, the new estimator, $\ddot{N}_h$, is

$$\ddot{N}_h = \frac{\left((n_{h_{1+}} - k_h) + 1\right)\left(n_{h_{+1}} + 1\right)}{\left(n_{h_{11}} - k_{h_{11}}\right) + 1} - 1. \tag{8}$$

As we are not trimming list B (CCS), $k_{+1} \equiv 0$, and therefore it does not need to be included.

Note, in the case $k_h = 0$, we return to the naïve DSE (equation 1) and, in the limit that $k_h = r_h$, $k_{h_{11}} = 0$, we would have the idealised estimator, Equation 7.

The only variables that we had available to us and that were appropriate related to the date of the most recent interaction with a data source.

There was no single date assigned to each individual that indicated a date of interaction that led to them being included in the SPD. Instead, we had to look at each data source individually and pull out the date of last interaction. Some data sources had multiple dates associated with them, such as CIS. Additionally, some of these dates referred to a period rather than a single date of interaction, such as tax year. See Annex 1 for a full list of variables available.

The process of trimming was further complicated by the fact that most people are not on every data source. This means we had to either extract a date from each source and score the data based on a single combined date, or score each data source and combine these scores. So far, we were more successful in implementing the second approach and is the one we kept for this paper.

We assigned to each record a score, with higher scores being more likely to be erroneous. These scores are related to percentiles of the population, so a score of 100 is assigned to those most likely to be overcoverage, and a lower score signifies a lower overcoverage probability.

To determine whether trimming was effective when it was applied at different score thresholds, we examined the impact on the estimates. If records were removed at random, $\frac{k_h}{n_{h_{1+}}} \approx \frac{k_{h_{11}}}{n_{h_{11}}}$, which would lead to $\widetilde{N}_h \approx \ddot{N}_h$, and trimming would only increase variance without reducing bias. If we removed more records that were erroneous than weren't, $\frac{k_h}{n_{h_{1+}}} > \frac{k_{h_{11}}}{n_{h_{11}}}$ and $\widetilde{N}_h > \ddot{N}_h$, therefore trimming was effective and overcoverage bias in the SPD was reduced. We continued trimming until the we reached the inflection point in our estimates shown in Figure 6.

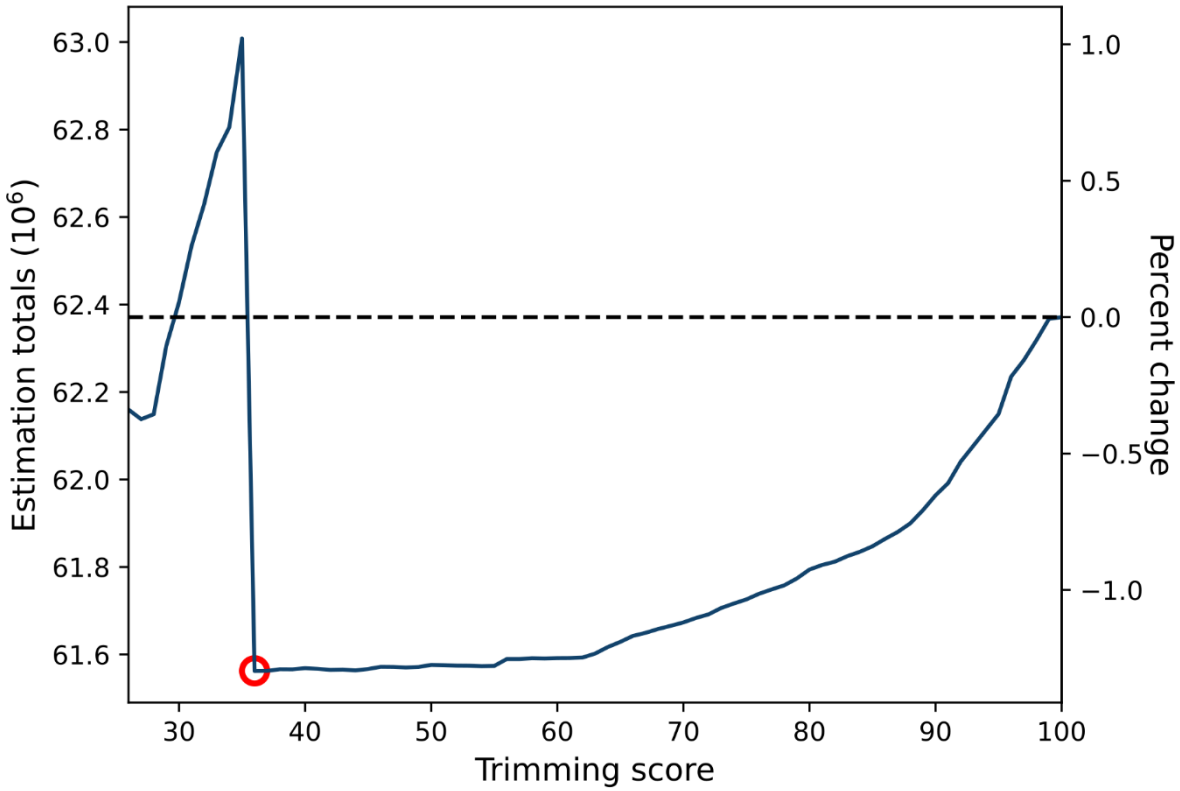


*Figure 6: Coverage-adjusted estimated national totals by trimming score. The red circle indicates the point at which we stopped the process (trimming score = 36).*

For many individuals, the only administrative records available were tax returns. As these records do not have a specific date, the end of the tax year was assigned as the latest interaction date with administrative sources. This led to many individuals having the same interaction date. Therefore, after the trimming threshold reached this date, a significant proportion of individuals were removed simultaneously, as can be seen in Figure 7. The trimming score corresponding to this date was used for the case study presented in this paper, and we expect trimming to become more effective when other administrative sources and variables are also taken into account.

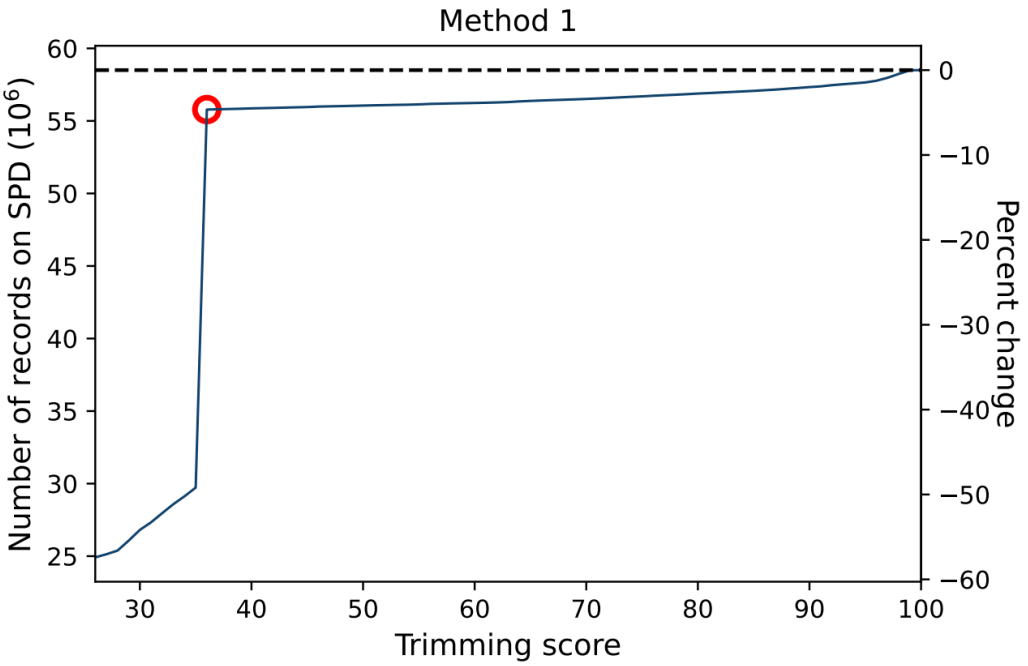


*Figure 7: Size of the trimmed SPD by trimming score. The red circle indicates the point at which we stopped the process (trimming score = 36).*

## Annex 4 – DSE, Ratio and Local Synthetic Estimation

As part of the 2021 SPD Case Study, we implemented the DSE method described in Annex 3, which estimates the population size using DSE post stratified by LA and 5 year age-sex groups across which included both CCS2 sampled areas and non CCS2 sampled areas. The method described in this section, mirrors the approach used in the 2011 Census of E&W. This method includes using the DSE, Ratio and Local Synthetic estimators to estimate the population sizes for LA by age-sex groups.

The DSE will be post-stratified by LA, hard-to-count (HtC), cluster of postcodes (OA) and 5 year age-sex groups, for CCS2 sampled areas only. Due to the small sample sizes of the CCS2 in the DSE, we will aim to either remove OAs from the sample where there are no CCS2 counts or collapse for strata where there are small or no CCS2 and SPD counts. The ratio estimator will then be used in combination with DSE to produce population size estimates for those not in the CCS2 sampled areas, for each Estimation Area (EA), HtC and age-sex group. Estimation Areas will be used here as in the 2011 Census, where LAs that had a small sample size were combined (Abbott, 2009, pp. 25 - 32). Collapsing LAs in this way reduced the risk of having small sample sizes to estimate the ratios. To estimate the required population size for each LA by age-sex group, the Local Synthetic estimator will be applied. Using the DSE, Ratio and Local Synthetic estimators is more desirable than just using the DSE to estimate the population size, as it allows us to deal with the heterogeneous response of individuals that vary by HtC and other characteristics by which response varies. The DSE-only approach also results in large variability of the

population size estimate because the coverage of the CCS2 will be low in the DSE, especially compared to the coverage of the SPD.

Dual System Estimation:

1) Subset the SPD population into CCS2 sampled areas. This will include individuals who responded to both the SPD and CCS2, those who responded only to the CCS2 and those who responded only to the SPD.
2) Due to the homogenous response probability assumption, stratify the population by Local Authority, HtC, output area and age-sex groups.
3) A contingency table will then be created for each of these strata, which will contain counts for individuals in both lists, and for those in only one list. There will be three observed counts for each of the contingency tables.
4) The overcount propensity will also be included here but will be at a higher level then described above due to small population sizes. The propensities will be applied to the contingency tables that are within the level defined for the overcount propensities. For example, if we estimated the overcount propensities by LA for each age-sex group, the propensity will be applied to the contingency table that sits within the specified LA and age-sex group.

$$\hat{t}_{alhc} = \frac{(\hat{g}_{as} X_{alhc} + 1)(Y_{alhc} + 1)}{(M_{alhc} + 1)} - 1$$

where $X_{alhc}$ is the SPD count for age-sex $a$ in LA $l$, HtC $h$, and cluster $c$;

$Y_{alhc}$ is the corresponding survey count;

$M_{alhc}$ is the corresponding SPD to survey match count; and

$\hat{g}_{as}$ is the corresponding overcount weight for LA supergroup $s$

Ratio Estimation:

1) Specify the strata that we want to estimate: EA by HtC by age-sex groups.
2) Sum up all DSE population size estimates in the sampled areas stratified by EA, HtC and age-sex groups.
3) Sum up all SPD counts in the sampled areas stratified by EA, HtC and age-sex groups.
4) Create the ratio between these estimates and counts stratified by EA, HtC and age-sex groups.
5) Sum up all SPD counts (inside and outside sampled areas) stratified by EA, HtC and age-sex groups.
6) Apply the ratios from (4) to the summed SPD counts in (5).

$$\hat{T}_{aeh} = \frac{\sum_{c \in S_{eh}} \hat{t}_{aehc}}{\sum_{c \in S_{eh}} X_{aehc}} X_{aeh}$$

$X_{aehc}$ – SPD count for age-sex *a*, in EA *e* HtC *h*, cluster c;

$\hat{t}_{aehc}$ – DSE estimate for age-sex *a*, in EA *e*, HtC *h*, cluster c ;

$X_{aeh}$ – SPD count for age-sex *a*, in EA *e*, HtC *h*

Local Synthetic Estimation:

To estimate the population size at LA by HtC by age-sex group level, the ratio between the estimated population size and observed count at EA by HtC by age-sex groups is applied to observed counts at LA by HtC by age-sex group level.

$$\hat{T}_{alh} = \frac{\hat{T}_{aeh}}{X_{aeh}} X_{alh}.$$

$X_{aeh}$ – SPD count for age-sex *a*, in EA *e* HtC *h*;

$\hat{T}_{aeh}$ – Population size estimate for age-sex *a*, in EA *e*, HtC *h*, cluster c ;

$X_{alh}$ – SPD count for age-sex *a*, in LA *l*, HtC *h*

## Annex 5 – Variance Estimation

Design-based and model-based approaches were considered for variance estimation, to estimate confidence intervals for an adjusted SPD 2021. We implemented the design-based approach. For this approach, given the stratified multistage sampling design of CCS, replicate (bootstrap) samples of PSUs were drawn separately within each design stratum (Wolter, 2007). This approach mirrors the 2021 Census variance estimation approach. The sampling design used for variance estimation mirrored the approach used to draw the 2021 CCS.

The CCS was stratified by LA and hard-to-count (HtC), with optimal allocation of OA and proportional (25%) selection of postcodes within each OA. The steps for implementation were as follows,

1. Start with the SPD linked to CCS2
2. Calculate the size of each stratum (LA by HtC), which is the number of OAs in each one.
3. Create a new count for each stratum of [the number of Output Areas – 1].
    a. Note: if the OA count = 1 then leave the count as it is.
4. Draw a pseudo sample (generated sample, not observed) of OAs (as well as the postcodes and observed individuals) from each LA by HtC using unrestricted random sampling with replacement. This means all OAs have equal probability of being selected within the stratum to which they belong.

5. Apply the DSE (with overcount propensities) to estimate the population sizes for each age-sex by LA stratum.
6. Repeat steps 4 and 5 for the required number of bootstrap samples (ideally 1000-2000).
7. Produce percentile confidence intervals of the population size estimates across all samples.

This approach takes into consideration the design of the 2021 CCS and allows for incorporation of different coverage adjustments, such as undercoverage, overcoverage, household size and bias adjustments.

During the quality assurance process of the methods and code, we were confident that our implementation was correct but our application and DSE stratification did not give the expected distributions centred on the point estimate. We are therefore reviewing whether the stratification used in estimation has caused this and whether the problem is solved by using the methods outlined in Annex 4. Therefore, we do not include any measures of uncertainty for the case study estimates we have produced.