Coverage estimation for admin-based population size estimation

Authors: Ceejay Hammond, Charlotte Hassell, Amy Large, Shuwei Lin, Mohammed Makhdoom and Alice White (ONS)

Contents

Key Messages of Paper

Purpose

- This paper provides an overview of our approaches to creating a coverage estimation process to support the production of admin-based population size estimates using the Statistical Population Dataset (SPD).

Recommendation

- Review post-stratification for the PECADO approach.
- Investigate the PECADO approach where tighter inclusion rules are used to reduce overcoverage error.
- Investigate alternative methods for population size estimation, such as multiple system estimation.

Key Asks of MARP

- Provide feedback and assurance on our approaches, conclusions, and next steps.
- Provide suggestions about any options we should investigate.

Introduction

The Administrative Based Population Estimates (ABPEs) require an unbiased stocks file as part of the input data for the Dynamic Population Model (DPM). The Statistical Population Dataset (SPD), which is the proposed source of that stock file, is susceptible to coverage issues and biases. As such, it is necessary to produce a coverage adjustment for that file, or to identify (and potentially coverage adjust) an alternative source.

Since the 2021 England and Wales Census, we have focussed on implementing methodology to produce this coverage adjustment to a defined set of quality standards. The paper outlines the steps that have been taken, taking 2021 data, concurrent administrative data and investigations into potential survey-based options into consideration.

Previous papers for MARP have covered methodological details. The purpose of this paper is to provide a holistic view of the process, as well as an update on current direction of travel and progress; acknowledging that work to date has not achieved the objective to the quality standards required.

We begin with an overview of the data estate we have been utilising for this work, then give summaries of the areas of focus to date and the results that have been observed, to demonstrate the evolution of the work and what has been tested so far. The final section covers what the current focus of the work is, in order to check with the Panel that this is an advisable next step.

Data

Statistical Population Dataset (SPD) V4.0

The Statistical Population Dataset is a unit level dataset which aims to approximate the usually resident population down to small areas with admin data. The aim of the SPD is to support the delivery of high-quality admin-based population estimates for the Dynamic Population Model (DPM). The SPD has a reference date which for SPD v4.0 is 30 June 2021 (ONS, 2023a). Records who are active within the year prior to the SPD reference date are included.

The SPD is subject to coverage error such as undercoverage error, particularly in older working ages, and overcoverage error (misplacement, duplication, and erroneous non-usual resident records), particularly in younger working ages.

Table 1: sources within SPD v4.0 (ONS, 2023b)

Name	Description
Personal Demographic Service (PDS)	National electronic database of NHS patient demographic details
English School Census (ESC)	Collection of pupil and school level data provided from all English local authorities for state schools only.
Welsh School Census (WSC)	Collection of pupil and school level data provided from all Welsh local authorities
Individual Learners Record (IRL)	Collects data on further education students and their studies from learning providers who receive funding from the Education and Skills Funding Agency
Higher Education Student Record (HESA)	Higher education data from across the UK.
Hospital Episode Statistics (HES)	Data covering individuals who have had admissions, outpatient appointments and A&E attendances at NHS hospitals in England
Emergency Care Data Set (ECDS)	Data covering individuals who have received treatment via urgent and emergency care (From 2021 ECDS replaces the A&E dataset HES)
Benefits and Income data (BIDS)	Provides data about specific benefits
Customer Information System (CIS)	DWP CIS is a database containing (almost) all residents of the UK who have a National Insurance Number (NINo).

Health Episode Statistics (HES)

The NHS's Health Episode Statistics dataset collects information about attendance, appointments and admissions to NHS hospitals in England. A key purpose of this data is for research and planning health services.

Between 2016 to 2019, HES combined three datasets: Admitted Patient Care (APC), Accident and Emergency (A&E) and Outpatients (OP). From 2021, Accident and Emergency was replaced by the Emergency Care Dataset (ONS, 2023b).

The data includes information about individuals' interactions with health services, key characteristics of the population, including ethnicity, so are an important source for administrative-based ethnicity statistics

Undercoverage error will exist in HES and ECDS as some individuals within the target population do not/rarely interact with healthcare services. For example, young adult males, students or migrants. Overcoverage error may exist in HES and ECDS as it may include records of individuals who have died or emigrated after interacting with the healthcare services (ONS, 2023b).

The ONS is supplied with data from NHS England both monthly and annually. The annual data is delivered to ONS in October, where the reference period for HES and ECDS begins in April. This results in an 18-month lag between the start of data collection for the annual supply and the data being available in the ONS. Therefore, the monthly data supply offsets the potential lag to support timelier data analysis (ONS, 2023b).

Personal Demographic Service (PDS)

The PDS is the national database of NHS patient demographic details of individuals who have interacted with an NHS service across England, Wales and the Isle of Man (ONS, 2023b). This includes GP practices and hospital visits, which highlights the relationship between PDS and HES.

Undercoverage error may exist in the PDS, for example individuals who have private healthcare or immigrants. Overcoverage error may exist in PDS, particularly in more urban local authorities. For example, individuals may interact with a healthcare service in one area and then move out of the area without deregistering (ONS, 2023b).

Driver and Vehicle Licensing Agency (DVLA)

For the 2021 Census, ONS was supplied DVLA data covering approximately 50 million registered drivers. For the purposes of coverage estimation, this was filtered to a specific period of activity (usually up to 1 year prior to the reference date of interest), based on 'update' behaviour. An update occurs when an individual renews their licence or provides new location (address) or personal (usually name) information to be included on their licence.

The DVLA data exhibits strong cyclic behaviours in renewals due to the 10-year life of a standard licence. This results in coverage peaks in the update file at ages (approximately) 18 – when many individuals acquire their first licence – and then at 10-year intervals. Once an adult reaches 70, this renewal window reduces to 3 years. There

are also shorter renewal windows for those with certain medical conditions and for those holding a higher category (bus or lorry) licence.

Due to this interaction pattern, the majority of behaviour is observed around these renewal points. Whilst there is a requirement to keep location and personal information up to date on your licence (including a fine if you do not inform of an address change), and there is no charge to update these details, unless it is directly observed (i.e. by the police), then these details being incorrect can easily go undetected, leading to misplacement and potentially leading to inaccurate linkage variables, damaging our ability to index the DVLA data against other sources. Similarly, if an individual is an international emigrant, this may also go undetected, leading to overcoverage error.

2021 Census responses

The online-first 2021 Census of England and was a compulsory usual resident population survey conducted on 21 March 2021 (reference date), collecting key characteristics and attributes of individuals and households. 97% of households responded across England and Wales and over 88% in all local authorities. 89% of returns were online (ONS, 2022b).

The 2021 Census has been indexed to the demographic index allowing us to use this as a coverage list within our approaches alongside the admin datasets (Shipsey, Law, Davies, Hammond, Pauna and Jones, 2024).

2021 Census Coverage Survey

The 2021 Census Coverage Survey (CCS) was used to measure coverage of the 2021 Census of E&W. This survey started eight weeks after Census day and was designed to produce an independent count from the Census for sampled areas across all local authorities in England and Wales. The sample contained approximately 16,000 postcodes, which is 1.45% of England and Wales postcodes (ONS, 2022a).

The CCS is an area-based sample with a two-stage cluster design. The CCS is stratified by local authority by hard to count index, where the sample is allocated to hard to count index by optimal allocation method and is then allocated to local authorities in proportion to their size (Burke and Račinskij, 2020).

We made use of a subset of the 2021 CCS, referred to as the CCS2. This is a 50% that approximately maintains the same level of representativeness of the CCS. This is due to the linkage between the 2021 CCS and demographic index, where the CCS2 subset was clerically linked and therefore satisfying the high-quality requirements of Census estimation (ONS, 2023c).

2021 Mid-Year estimates (MYEs)

The Census-based 2021 MYEs provide a benchmark for us to compare our population size estimates to. We assume the MYEs are the 'truth' and therefore we can calculate coverage error (%).

Approaches and results – timeline

Research has been carried out to explore the feasibility of producing more timely estimates of population and population change. As part of this work there is a requirement to provide high quality, approximately unbiased population stock estimates as one of the inputs to the Dynamic Population Model (DPM). For this purpose, a coverage estimation process is currently under development to provide an adjustment to the Statistical Population Dataset (SPD). Since beginning this work, we have investigated and implement both admin-survey based approaches and admin only based approaches to population size estimation. Below we discuss the approaches (in order of implementation), methods, results and conclusions which have informed our recommended next steps.

2021 CCS Case Study

This section will give a summary of the work we did for the 2021 CCS case study. Law, Large, Hammond and Linton (2023), present a more detailed overview with key results.

This case study was an initial attempt at an admin-based population size approach. Our aim was to understand the potential of applying traditional census methods, the dual system estimator (DSE) and overcount propensity (Račinskij, 2018 and Large, Brown Abbott and Taylor, 2011) to admin data to estimate the population size and to create an unbiased estimate of the population size by estimating undercoverage and overcoverage error within the SPD v4.

This approach aimed to produce a coverage weight for the 2021 SPD v4 by age, sex and local authority (LA). The data used in this approach made use of high-quality linkage available between the SPD v4 and the CCS2. As the CCS does not collect information on Large Communal Establishments (LCEs, more than 49 bed spaces), we estimated coverage of the SPD for households and small communal establishments (SCEs, 7 to 49 bed spaces) and used 2021 census LCE estimates to be able to compare to the 2021 MYEs.

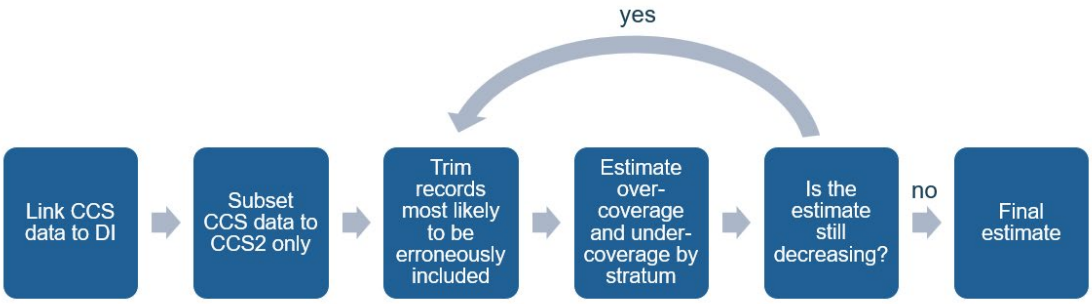
Dual System Estimator Key Assumptions:

1) The population is closed (for example, no births or deaths).
2) Perfect matching between the two lists.
3) The capture/inclusion probability of individuals in at least one of the lists is homogeneous.
4) The individual's inclusion probabilities in the lists are independent.
5) There are no erroneous captures in the lists (overcoverage).

Undercoverage error was estimated by post-stratifying the SPD and CCS2 by age, sex, LA and hard-to-count and implementing the dual system estimator. To estimate overcoverage error (misplacement, duplication and non-usual residents), the overcount propensity was used to downweigh the population dataset (SPD v4), where the population was post stratified by age-sex group and LA supergroups. To support the overcount propensity, CCS2 sampling weights were also used to correct for the higher probability of inclusion of some postcodes in the CCS2. An initial trimming approach was used to remove overcount (erroneous records) from the SPD, by making use of interaction date and income.

The estimation process was 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.
3. Add LCE totals to estimated totals for private households and SCEs for comparison to MYEs



Comparing the coverage adjusted SPD v4 to the 2021 MYEs at national level resulted in 3.98% coverage error, whereas at subnational level, some LAs were overestimated by over 15% compared to the 2021 MYEs. We concluded that although we deemed these results unusable, they provided us with a useful indication of the quality that may be needed with similar types of data and methods in the future. We proposed improvements to this method included trimming to remove erroneous records from the population dataset (list A) and a calibration approach to estimate misplacement error within list A.

Survey Simulations

We investigated an approach to population size estimation where the population dataset is subject to undercoverage error only and overcoverage error is negligible. We made use of an area-based sample survey to estimate undercoverage error, and overcoverage error is reduced in the population dataset to a negligible level by using strict inclusion rules or model-based trimming scores.

For these simulations 2021 Census responses was used as the population dataset, where samples, coverage error, trimming rate, and individual inclusion/response to the population dataset and sample were specified and generated. To estimate the population size at national and subnational levels, we implement the DSE and overcount propensity (Annex 1). The target population for the simulation study consisted of usual residents in households only and did not include communal establishments. Sample design included a two-stage cluster design which mirrored the 2021 CCS sample design (Burke and Račinskij, 2020) and systematic sampling, used for the Labour Force Survey (LFS) (ONS, 2024).Limitations of the simulations:

- It is difficult to mirror everything we would observe in the admin data/ survey.
- We assume perfect linkage between the lists.
- Large numbers of simulated runs are difficult to implement due to processing time and memory issues.
- The simulations did not use the most up to date version of the trimming methodology and therefore does not present the most recent modelling inclusion models.

Results showed that under the specified quality requirements, the sample size required to meet the variance criteria was substantially larger than that needed for the bias criteria. The size needed to meet the criteria at a Local Authority level was also larger than could realistically be achieved given operational constraints and challenges with falling response rates for voluntary survey activity. For a survey to become a viable option, a substantially different design (for example collecting smaller samples over a number of years and combining them over time) and different approaches to data collection would need to be considered, which was not within the scope of the project at that time.

PECADO approach

In parallel to the survey simulations, we began to explore an admin-based only coverage estimation approach. An approach has been developed by Dunne and Zhang (2023) to estimate the population of the Republic of Ireland, without the use of a purposefully designed coverage survey or Census. The approach is referred to as the Population Estimates Compiled from Administrative Data Only (PECADO).

The Ireland PECADO approach makes use of two lists for estimating the population:

1) A population dataset constructed through linkage of administrative data sources (list A)
2) Driving Licence Data (DLD) (list B)

To estimate the population size, the trimmed dual system estimator (where the population dataset is trimmed of erroneous records) is implemented making use of the two lists. The DLD data was chosen as the second list as it is assumed there is negligible overcoverage.

We have attempted to mirror the PECADO approach to estimate the population for England only, where the population was post-stratified by 5-year age group, sex and local authorities (LA) across England using the Calibrated DSE (cDSE) and the Calibrated Trimmed DSE (ctDSE). The results do vary by age-sex-LA, but for ease of discussion, only the aggregated age-sex (England level) results are presented here.

Methodology

The proposed methodology was developed by (Zhang, 2023). We implemented both the ctDSE, which incorporates the dual system estimator (DSE), trimming the population dataset of erroneous inclusions and a calibration for the misplacement of individuals within the population, whereas the cDSE does not include trimming.

Both the ctDSE and cDSE is used to estimate the total population size for domain i, where misplacement at subnational levels is present in the population dataset. We estimate the level of misplacements, where individuals are not in the correct location at sub-national level.

The calibration is calculated by making use of the trimmed population dataset and estimated inclusion probabilities, the observed misplacement counts and the ratio of the number of individuals who matched between the two lists in the same and different locations. The calibration is then applied to the observed level of misplacement from linkage between the two lists, where list B is assumed to determine an individual's correct location.

Once the calibration component has been estimated and therefore the level of misplacement in domain i, this is then removed from the trimmed population dataset count for that domain.

Therefore, to estimate the total population size $\widehat{N}_i$ without calibration, the Trimmed DSE (**tDSE**) is denoted by,

$$\widehat{N}_i = \frac{n_i * x_i^*}{m_i^*}$$

To estimate the total population size $\widetilde{N}_i$ where calibration is applied, the Calibrated Trimmed DSE (**ctDSE**) is denoted by,

$$\widetilde{N}_i = \frac{n_i * (x_i^* - \hat{\alpha}_i)}{m_i^*}$$

Where,

x^* = total number of records in the trimmed population dataset
n = total number of records in list B
m^* = total number of matched records between the trimmed population dataset and list B

$\hat{\alpha}$ = the calibrated estimated number of misplaced records in the trimmed population dataset

Modelling Inclusion

The aim of this work was to remove (trim) the overcoverage records from the SPD using inclusion models. These records are individuals who are not in the usual resident population of England and Wales but were included in the SPD due to their interaction with a service. The inclusion models predict a probability of each SPD record corresponding to a usual resident. Two approaches to trimming with the predicted 'inclusion scores' have been considered throughout our work, and a more detailed discussion is presented by Shipsey et al., (2024).

- Automatic trimming:

  This was the initial method we used to support our estimation approaches. This method uses list B in the ctDSE to determine when trimming becomes less effective and therefore is dependent on which list B is used in estimation. This is achieved by ordering records by inclusion score from low to high and finding the point at which further trimming removes linked cases at a similar rate to unlinked cases.

- Quality-tuned trimming:

  This is the method we have most recently used and is used for the results in this paper. This method does not depend on list B and instead assumes the probability scores derived from the model are correctly calibrated. Probabilities output by inclusion models are not calibrated to the true probability of inclusion in the population, but instead the lower probability of being linked to a census response. Therefore, the modelled probabilities are adjusted to account for non-response to the Census.

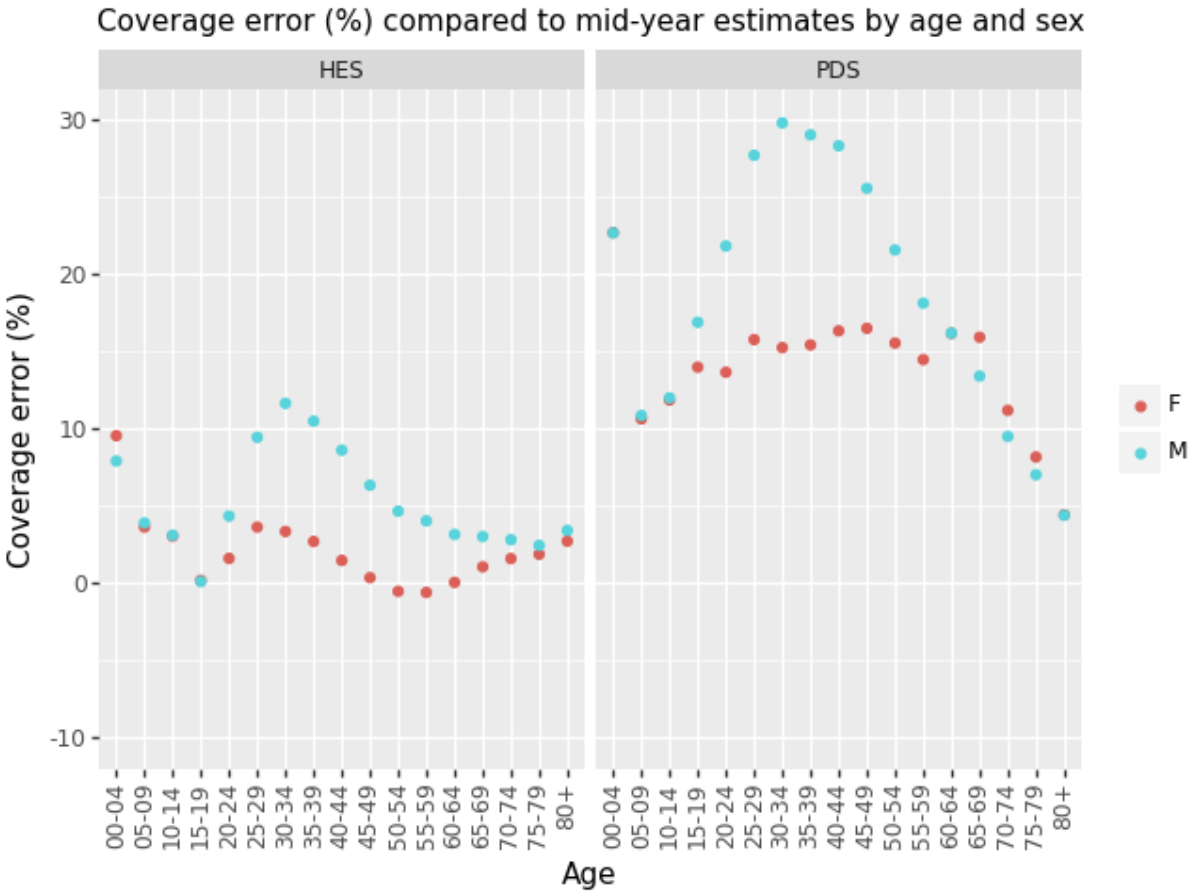
Results

Table 1: PECADO scenarios

Scenario	List A	List B	Method
1	SPD excluding HES	HES	cDSE
2	SPD excluding HES	2021 Census responses	cDSE
3	SPD excluding HES & PDS	HES	cDSE
4	SPD excluding HES & PDS	HES	ctDSE
5	SPD excluding HES & PDS	PDS	cDSE
6	SPD excluding HES & PDS	2021 Census responses	cDSE
7	SPD excluding HES & PDS	2021 Census responses	ctDSE
8	SPD	DVLA (1 year)	cDSE
9	SPD	2021 Census responses	cDSE
10	SPD	2021 Census responses	ctDSE
11	2021 Census responses	HES	cDSE

Results from the PECADO approach implementations are presented below for scenarios 2, 3, 6, 7, 8, 9, 10 and 11. These results present applications of both the cDSE and ctDSE with a combination of lists. Coverage error (%) is calculated between the estimates and the 2021 MYE, where the 2021 MYEs are assumed to be the 'true' population size. We have not presented the results for scenarios 1 and 2 as these scenarios involve producing list A by filtering out only HES from the SPD. For scenario 1, as an individual's probability of inclusion in HES and PDS is dependent on the other list, not filtering the SPD by both HES and PDS will result in violating a key assumption of the DSE.

Figure 1: List A = SPD excluding HES + PDS, List B = HES or PDS



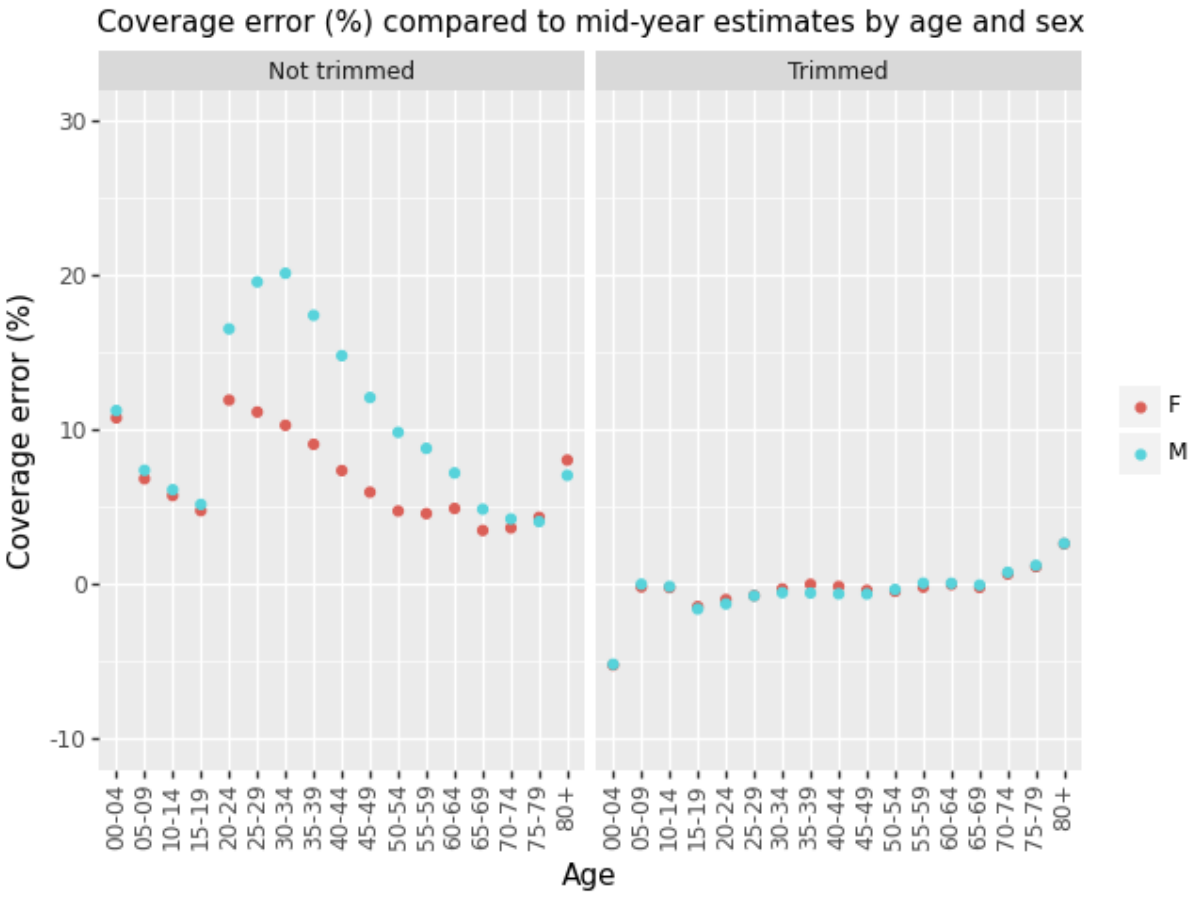
These results compare the implementation of the cDSE where List A remains the same across both scenarios and list B differs between HES and PDS, both of which are health datasets (corresponding to scenarios 3 and 5 in table 1).

When list B is HES, across almost all age-sex groups the coverage error (%) is positive. The age groups with the highest coverage error are 0–4-year-old males and females, and 25–49-year-old males. Apart from the youngest and oldest ages, males have higher coverage error (%) compared to females. We expect this is due to overcoverage error among these age-sex groups, which would result in an overestimation of the population size. For some age-sex groups such as 15–19-year-old males and females, and 50 – 64-year-old females coverage error (%) is close to zero. This suggests this scenario performs well for some age-sex groups, or combined errors within the lists may cancel out.

When list B is PDS, the coverage error (%) across all age-sex groups is positive. Particularly for 0–4-year-old males and females, and 20–55-year-old males. Apart from the youngest and oldest ages, males have higher coverage error (%) compared to females, where females have a constant coverage error (%) of around 15% between 15–69-year-olds. We expect this high coverage error (%) comes from overcoverage error in list A and list B, which would result in an overestimation of the population size. Even for the age-sex groups with the lowest coverage error (%), 5-19 and 70+, the coverage error

(%) is still large. This shows this scenario does not perform well for any of the age-sex groups.

Figure 2: List A = SPD, List B = 2021 Census responses



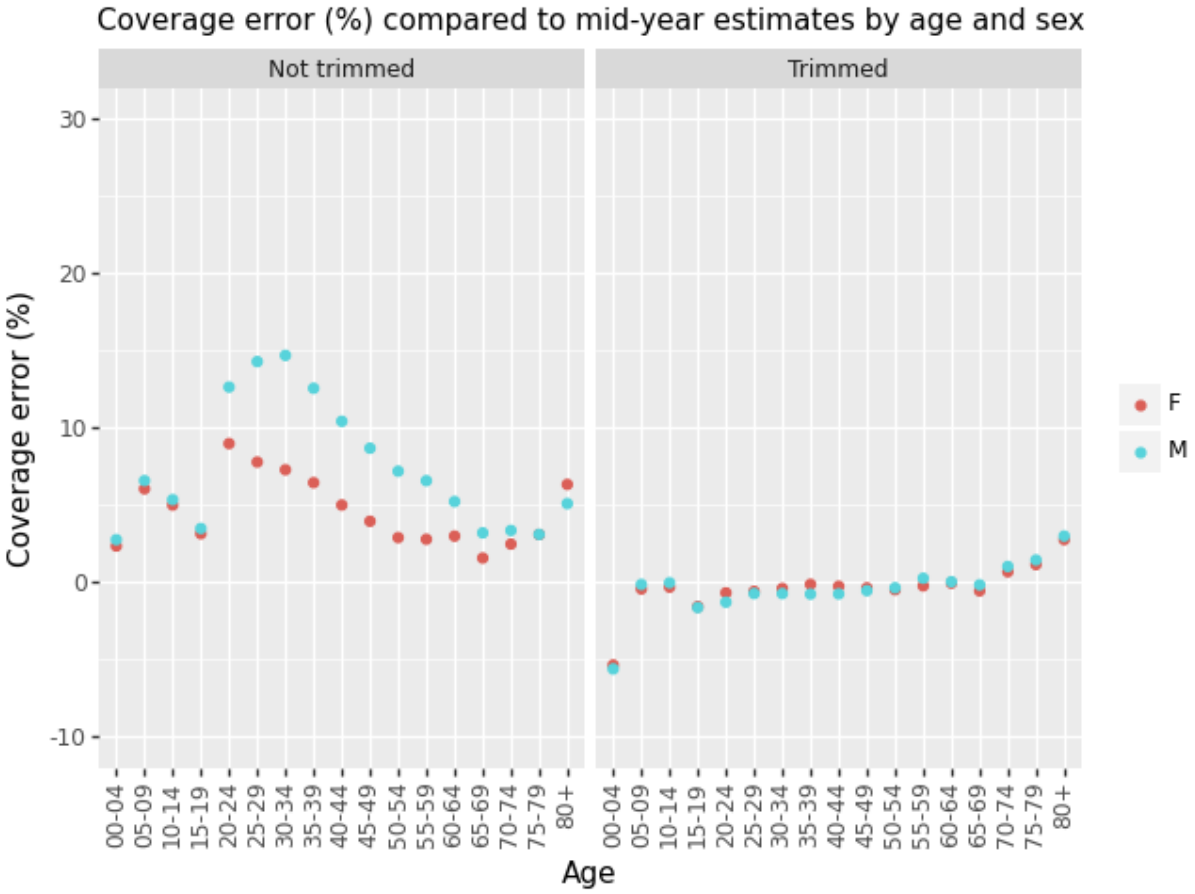
These results compare the implementation of the cDSE and ctDSE where List A and List B remains the same across both scenarios to understand the impact of trimming (scenarios 9 and 10 in table 1). We used 2021 Census responses as list B here to understand and identify bias in the admin data only scenarios.

The ‘not trimmed’ results show large positive coverage error (%) across all age-sex groups, indicating that this scenario does not appear to be appropriate for any of the groups. Similar to previous discussed results, coverage error (%) is similar for males and females for the youngest and oldest ages, however between 20-64, males have larger positive coverage error (%) compared to females. Mirroring previous conclusions, we believe this most of this error may come from overcoverage error in list A, which results in overestimation of the population size. Linkage failure could also be contributing to the error.

The ’trimmed’ results support our previous conclusions. When list A is trimmed of erroneous records, using the quality-tuned approach with the GBT_NRM1 model, the

coverage error (%) across all age-sex groups, apart from 0–4-year-olds is around 0. Although the results look promising, there are limitations of this quality-tuned approach, such as its reliance on a well calibrated model using 2021 Census data and the effects of data drift and model drift over time will gradually affect the accuracy of the model's predicted probabilities and therefore would require periodic calibration, for example through the use of a coverage survey (Shipsey et al., 2024). A further complication is in finding a suitable administrative list which would emulate all of the desirable properties we have in the 2021 Census responses.

Figure 3: List A = SPD excluding HES & PDS, List B = 2021 Census responses

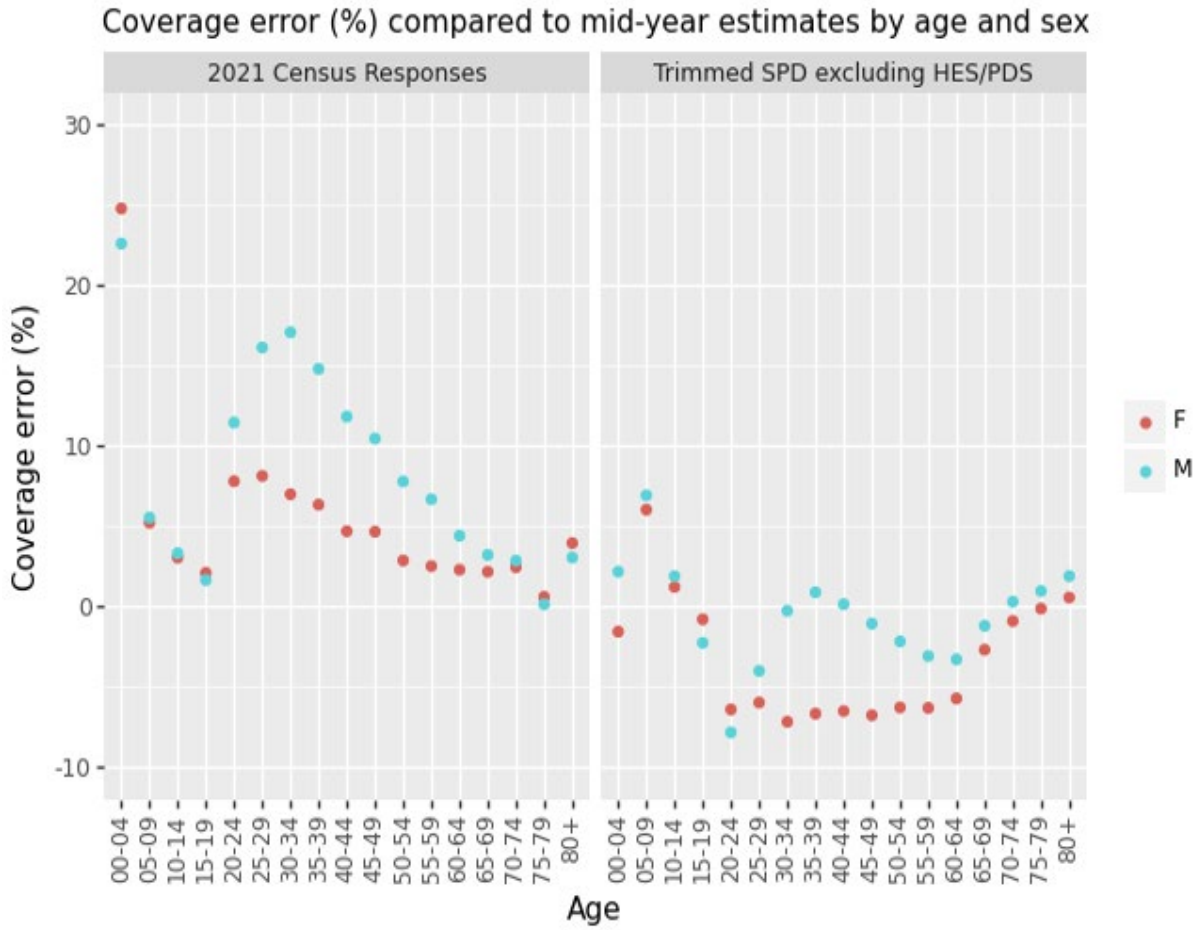


These results compare the implementation of the cDSE and ctDSE where List A and List B remains the same across both scenarios to understand the impact of trimming (corresponding to scenarios 6 and 7 in table 1)].

The results in Figure 3 mirror the results presented in Figure 2, showing there has been minimal impact from removing the health sources from the SPD when used alongside such high quality list B. The 'trimmed' results show the impacts of using the GBT_NRM1 model. Trimming does reduce coverage error (%) across age-sex groups and results in

some negative coverage error (%), particularly for 0–4-year-olds. This scenario has the same limitations as those for the results in the previous figure.

Figure 4: List A = SPD excluding HES + PDS, List B = HES and List A = 2021 Census responses, List B = HES

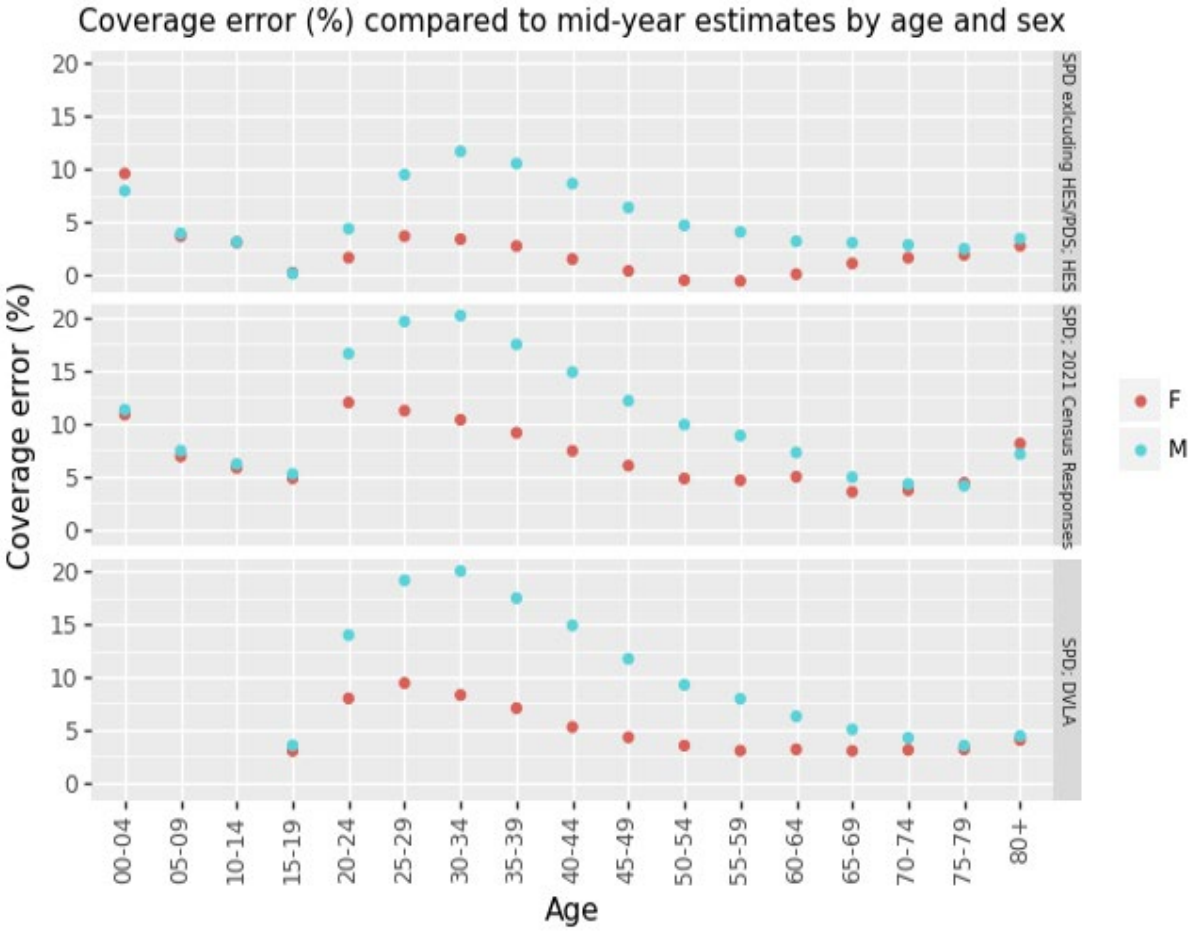


These results compare the implementation of the cDSE and ctDSE where List A is either SPD excluding HES + PDS, which was trimmed using the quality-tuned approach (GBT_NRM1 model) due to erroneous records or 2021 Census responses where no trimming was applied due to negligible overcoverage in the Census responses. List B remains the same across both scenarios to understand the impact of different List As (corresponding to scenarios 4 and 11 in table 1).

When list A is 2021 census responses coverage error (%) is positive across all age-sex groups. Groups with highest coverage error (%) include 0–4-year-old males and females and 20–49-year-old males. Similar to previous results, the largest difference between males and females are between 20–64-year-olds. This source of the positive coverage error (%) may come from overcoverage error in list B or linkage error between the two lists.

When list A is the trimmed SPD excluding HES + PDS, coverage error (%) is negative across the majority of the age-sex groups. Groups with highest coverage error (%) include 5–9-year-olds and 20–49-year-old females. These results suggest that whilst the trimming is reducing the level of overcoverage in the data, it is also highlighting or even causing a failure of the homogenous capture assumption for DSE.

Figure 5: List A = SPD excluding HES + PDS, List B = HES and List A = SPD, List B = 2021 Census responses and List A = SPD, List B = DVLA



*Note: DVLA results start with **16-19** age group, not 15-19.*

These results compare the implementation of the cDSE across three scenarios to understand the impact of different List As and list Bs (corresponding to scenarios 4, 8 and 9 from table 1).

For the scenario where the SPD is list A and DVLA is list B, results are for 16+ year olds, due to small counts for 15-year-olds and no inclusion of younger ages in the DVLA. All three scenarios exhibit similar trends in terms of coverage error by age and sex, but the scenario using HES as list B gives the lowest overall coverage error.

The difference between the results can be explained by overcoverage error within the lists or other sources of error. Males generally have higher coverage error (%) than females.

Looking at all the scenarios produced to date, the combination providing the most successful national level results (by age and sex) only using administrative data is the SPD excluding HES & PDS as list A and HES as list B.

Data Deep Dive

The data deep dive began as an exercise to investigate causes for biased results when using combinations of available administrative data in using the PECADO methodology, detailed above. It was believed that a failure of a DSE assumption, or an error in processing and implementation of the method was a likely cause for the results. The deep dive was therefore tasked with looking at the full life cycle (from data creation to processing and production of estimates) of the data sources being used. The following areas were identified as possible sources of error / assumption failure:

- We have used the right data (processing)
- Our code is error free (processing)
- These is no overcoverage in either list (DSE assumption)
- Linkage is perfect (DSE assumption)
- There are no excluded populations, so every member of the target population has a non-zero probability of being on one of the lists (DSE assumption)
- Our target is the usual resident population of a defined geography (DSE assumption)
- The reference date is the mid-year point, and the population is closed (DSE assumption)
- We have structural independence of our lists - inclusion (or not) on one list is not conditional on inclusion (or not) on the other (DSE assumption)
- We have homogeneity - probability of capture is similar across the population, with stratum
- Stratification variables are defined consistently and accurately on both sources

The deep dive data exercise was unable to identify a single explanation for the persistent bias in the estimates produced via the PECADO DSE method, using the SPD, DVLA data and various health sources.

The exercise did, however, demonstrate the following key messages:

- In terms of data creation and delivery there is a gap between those creating that data (i.e. data supplier side) and those using the data within ONS. Those directly communicating are often those involved in an operational role, and not the analysts

who have the detailed knowledge. Going forward, this relationship could be strengthened to get better clarity around data issues more quickly.

- The DI is the central construction for all our key administrative data sources. There is ongoing quality assurance of the linkage within the DI underway (report due in December 2025). This includes the impact of potential false negative links (where linkage protocols are stringent) as well as false positive ones, and how this subsequently impacts both the SPD construction and the indexing of any additional sources that could be used for an independent coverage list.
- Key population groups are consistently poorly represented in the data we currently hold and are very difficult to classify and identify as persons within these groups when we do find them. Differential response patterns are closely linked to the function and purpose of the administrative data. (For example, those living in large urban areas with good public transport may not hold a driving licence.) Some of the groups of concern include:
    - Migrants – this could be improved by the potential use of home office migration data, but this is more likely to have value as part of a trimming or modelling inclusion process to remove these individuals from the datasets being used. We will get good representation of those requiring visas to work and travel, but very little indication for those who do not require visas and for UK resident out migration.
    - Communal establishment populations – we know these individuals are in the data, but due to lags, the pattern of movement to and from CEs and proxy reporting, we are unlikely to get a clear picture of this group. If communal establishment residents require a different coverage approach to the household population, it will be difficult to isolate and/or remove these individuals from the data to allow such a treatment.
    - Other special population groups – there is ongoing work looking at how groups such as the traveler community, the homeless and certain community groups appear in the data. This will need to be assessed in an ongoing manner.
- Whilst quality checks throughout the ingest of the admin data, and all subsequent processes (such as DI construction and SPD build) are carried out, these are owned across different parts of the process and, as such, the results are not brought together to provide a cohesive and holistic view of the quality of the data at every stage, and for every product created. The result is a system that is challenging to quality assure as a whole, with results for these assessments being held in multiple locations and inconsistently reported. While each check brings new information and understanding to the data, bringing it together in a more structured way would support the understanding of the overall quality of data assets and any potential limitations for subsequent usage.

- There was no conclusive evidence of specific failures in the DSE assumptions for the PECADO approach. Looking through an attributes lens reinforced existing understanding about presence and consistency within the data sources of interest. We therefore conclude that it is not a single point of failure in the assumptions of the DSE, but failures in multiple areas that are contributing towards the biased results.

Next steps: Multiple system estimation

We propose to implement a multiple system estimation (MSE) approach developed by Li-Chun Zhang which allows for estimation of both undercoverage and overcoverage error (Zhang, 2015). In this approach, three lists are discussed. List A and list B, both admin datasets with undercoverage and overcoverage error. List S, a survey with undercoverage error only.

This modelling approach is an extension of the Capture Recapture (Wolter, 1986) model underlying the dual system estimator (DSE) used for census undercoverage estimation, where overcoverage error is included. Li-Chun Zhang makes use of standard loglinear models, where an approach based on psuedoconditional independence is examined.

Li-Chun Zhang (2015) proposes two models which use the outcomes from the linkage between list A, list B and list S to estimate the error rate for observed outcomes. These are referred to in his paper as model (10) and model (11).

The aim of implementing this approach with the current chosen lists is a proof of concept and for us to understand underlying assumptions and methodology. Limitations of this approach include,

1. Does there exist a suitable list S to support this method?
2. For the assumptions of erroneous records underlying models (10) and (11), can we find a suitable 'high quality' list A and list B? or,
3. Can we suitably trim list A and list B?

We propose next steps for the initial implementation of this approach,

1. Initial implementation: List A = SPD excl HES & PDS, List B = HES and List S = 2021 Census responses
2. An approach with trimming of list A and list B (using already produced trimming models): List A = SPD excl HES & PDS, List B = HES and List S = 2021 Census responses
3. An approach with trimming of list A and list B (using updated trimming model for A and tighter inclusion rules for B): List A = SPD excl HES & PDS, List B = HES and List S = 2021 Census responses
4. An approach with trimming of list A and list B (using updated trimming models): List A = SPD excl HES & PDS, List B = HES and List S = 2021 Census responses.

If initial results look promising and give assurance, we can make use of the survey simulations and design to investigate a survey that could be a suitable candidate list S.

Discussion

The initial scope of this work was to determine if it were possible to produce a coverage estimation approach using administrative data. Throughout the work, we have encountered challenges with differing levels and patterns of both overcoverage and undercoverage on the sources (and source combinations) that we have used. The work has also been restricted to focussing on 2021 data to try and provide a suitably high quality data point (i.e. the 2021 Census estimates) to compare results to.

The work to date has indicated that we have not found a suitable combination of administrative data sources to provide us with the quality of estimates that we require as input to the DPM. We have also not yet achieved an accurate enough trimming model to remove all of the overcoverage from our sources. We have not been able to answer the question of why our methods are struggling with the combinations of data we are using, nor what the ideal set of attributes we are looking for in an admin source to make our methods viable. It is likely that challenges are arising from a compilation of errors that are accumulating from data production and maintenance, linkage, processing and manipulation for estimation.

We propose to implement more complex methodology, namely MSE, which will allow us to estimate for some of these compounding errors, such as residual overcoverage, whilst continuing to investigate alternative data combinations and the evolving data estate.

The ask, at this point, is do we feel there are any paths of investigation we have not looked at that we should have? And is our move to MSE a sensible one?

References

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

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).

Large, A., Brown, J., Abbott, O. and Taylor, A., 2011. Estimating and Correcting for Over-count in the 2011 Census. Survey Methodology Bulletin, 69, pp.35-48

Law, A., Large, A., Hammond, C. & Linton, M., 2023. Population stock estimates using linked administrative data and a coverage survey – a case study for 2021 and future directions. [Online] Available at: EAP202_Population_stock_estimates_case_study_for_2021_and_future_directions.pdf

Office for National Statistics (ONS), released 9 November 2022(a), ONS website, methodology article, Coverage estimation for Census 2021 in England and Wales

Office for National Statistics (ONS), released 2 November 2022(b), ONS website, statistical bulletin, Population and household estimates, England and Wales: Census 2021, unrounded data

Office for National Statistics (ONS), released 28 February 2023(a), ONS website, article, Developing Statistical Population Datasets, England and Wales: 2021

Office for National Statistics (ONS), released 3 March 2023(b), ONS website, article, Administrative sources used to develop the Statistical Population Dataset for England and Wales: 2016 to 2021

Office for National Statistics (ONS), released 27 June 2023(c), ONS website, methodology, Dynamic population model, improvements to data sources and methodology for local authorities, England and Wales: 2021 to 2022

Office for National Statistics (ONS), released 27 August 2024, ONS website, quality and methodology information report, Labour Force Survey (LFS) QMI.

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

Shipsey, I., Law, E., Davies, S., Hammond, C., Pauna, H. & Jones, S., 2024. Modelling inclusion to trim over-coverage and improve DSE for population estimation. s.l.:Internal, available upon request.

Wolter, K. 1986. “Some Coverage Error Models for Census Data.” Journal of the American Statistical Association 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

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

Zhang, L.-C. (2015). On Modelling Register Coverage Errors. *Journal of Official Statistics*, *31*(3), 381-396. https://doi.org/10.1515/jos-2015-0023

Annex 1: Simulations coverage estimation methodology steps

1) Two lists (admin-based population dataset and coverage survey) are linked.
2) Post-stratify the population by Local Authority, hard to count and age-sex groups (I).
3) For each strata a two-way table will be created, which will contain counts for three observed outcomes individuals in both lists, and in either list only. There is one un-observed outcome which are those who are not in either list.
4) The overcount propensity will also be included, however will be post-stratified at a higher level (L) 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.
5) To estimate the population size for each strata, use the Chapman corrected dual system estimator, where the overcount propensity is estimated and used to adjust the list A count.

$$\widehat{N}_i = \frac{(\frac{x_i}{\gamma_l} + 1) * (n_i + 1)}{(m_i + 1)} - 1$$

x = total number of records in the population dataset
n = total number of records in list B
m = total number of matched records between the population dataset and list B
γ = overcount propensity