SPD Estimation Options

Eleanor Law, Ceejay Hammond, Mark Linton, Zainab Ismail, Shaun Davies, Isaac Shipsey

September 2022

Key Messages of Paper

Purpose

In this paper we draw together proposed methods and data sources for population estimation using Statistical Population Datasets (SPDs) and without a traditional census[1]. We suggest some possible approaches that combine these data sources and comment on their benefits and limitations to inform the 2023 Recommendation.

Recommendation

We describe some combinations of data and methods that could be used, ranging from the most easily applicable now to the most sustainable and efficient. We also outline further research using simulations and tests with real data to find out the expected quality that may be achievable.

Key asks of MARP

We would like feedback on the approaches we have outlined, suggestions of any other ways to combine data and methods, and especially any comments on the kind of coverage problems and issues that the proposed methods will not address. We have placed some questions in boxes but comments on any part of the paper will be useful.

1. Executive Summary

One consideration of the 2023 Recommendation is whether there is a future need for a decennial census. If there is not, population estimates will need to be compiled from administrative sources supported by smaller scale survey data collection where required. ONS intends to produce coherent stocks and flows estimates for population and migration using the dynamic population model (DPM). Many sources will input into the DPM, but the main source for population stocks will be SPDs. For the DPM to produce reliable and unbiased population estimates, it requires stocks estimates that are unbiased, together with measures of uncertainty.

To secure the future production of trusted population estimates from the DPM, raw counts from SPDs should not be used as inputs to the DPM without quantitative information about error in the SPDs. Coverage errors should be inputted alongside SPD counts, or SPDs should first be corrected by using estimation methods to calculate appropriate factors that can be applied to the SPD. The aim of the estimation methods we describe in this paper is to provide these factors, or weights,

[1] That is, in its current decennial form.

at record level or aggregate level, to meet the requirement of the DPM to estimate single year of age by sex by Local Authority (LA) with appropriate measures of variance.

We summarise in this paper the main administrative and survey data sources currently available. We also summarise the estimation methods that have been proposed to bring together these sources, with their strengths and limitations. We briefly describe the kind of simulations we are using to build on previous work, and we share some examples of the type of output we have started to produce to assist with choosing the most appropriate methods.

We outline three high-level options for population estimation, which have different requirements for estimation of under-coverage and over-coverage. With the data available now, it is possible to use a combination of data sources to estimate both under-coverage and over-coverage of SPDs (Option 1). A more sustainable option in terms of cost, while providing high quality estimates, is to use a population or address dataset with negligible under-coverage so that an under-coverage survey is not required (Option 2). We describe how data sources and methods can be combined to deliver these two options and a third option of using a central population register that we believe is less realistic in the UK.

There are many challenges that remain and are not discussed in this paper in any detail, including consideration of communal establishments, specific questions that should be asked in coverage surveys, and the impact of linkage quality on estimation. We describe the future research work that we believe is required before making final decisions on the viability of producing population estimates with acceptable bias and variance from SPDs.

2. Introduction

Statistical Population Datasets

Statistical Population Datasets (SPDs), formerly known as Admin Based Population Estimates (ABPEs) are a mid-year approximation of the usually resident population of England and Wales using administrative data records **(ONS, 2021a)**. SPDs, as counts of population stocks, are intended to be one of the inputs to the Dynamic Population Model (DPM) from which ONS intends to produce demographic accounts in the proposed transformed system of population statistics. The DPM requires an ongoing source of unbiased population stock estimates. In the context of the National Statisticians 2023 Recommendation on the future of the census for England and Wales, we must be able to demonstrate our ability to produce unbiased estimates with a known level of uncertainty from other non-census sources. SPDs are currently the preferred predominant source but may be supported by survey sources where necessary.

ONS has been producing SPDs for several years and has iteratively developed the methods in response to quality evaluations undertaken against official population estimates. While administrative data are typically available at an increased frequency compared to survey data, these data sources are not collected for the purpose of

producing population statistics and can vary in their coverage, collection processes and variable definitions. SPDs are constructed by combining a range of sources (for example Department for Work and Pensions and/or Higher Education Statistics Agency data) to attempt to create one consistent picture of the population, which has good coverage by age, sex, and LA (ONS, 2013). They are produced by using the Demographic Index (DI) component of the ONS Reference Data Management Framework (RDMF) to provide a list of potential candidates for inclusion and then applying a set of deterministic rules to filter those who are not likely to be part of the usually resident population in any given year. A usual resident of the UK is anyone who, on a given reference day, is in the UK and has stayed or intends to stay in the UK for a period of 12 months or more, or has a permanent UK address and is outside the UK and intends to be outside the UK for less than 12 months.

SPDs can be aggregated to produce population counts, and **methods have been developed to indicate a kind of uncertainty** **(ONS, 2020a)** from variations between these raw counts and Census 2011. This method will soon also be applied to ONS Census for England and Wales 2021 (Census 2021). However, the uncertainty measures do not come from a sample so, in their current form, SPDs are not considered formal statistical estimates of population stocks. Previously, work has been carried out within ONS Methods and Quality Directorate (MQD) to explore methods of creating population estimates from SPDs. This work has assumed the **use of a large population coverage survey** **(ONS, 2011b)** known as the "**Integrated Population and Characteristics Survey**" (IPACS), **(ONS, 2019)** one component of which will take the form of the Labour Market Survey (LMS). The use of an ongoing survey will be essential for any methods that will estimate the under-coverage and over-coverage of the SPDs. The only alternative that may be considered is a reliable administrative source (or sources) independent of SPDs that can be assumed to have no over-coverage. Currently, we have no viable options for administrative data sources that could be delivered regularly and include negligible over-coverage. We include DVLA data in our table of available data sources to describe its limitations (Annex 2, Table A1).

Coverage of Statistical Population Datasets

Even an exceptionally well designed and implemented census of a large and complex population cannot provide an enumeration with ignorable coverage errors **(Račinskij, 2018)**. To a greater extent, this is also the case for counts based on population datasets constructed from administrative data.

The census is designed to estimate the population, whilst administrative datasets are not produced for this purpose. Administrative datasets are usually collected by governments or other organizations for non-statistical reasons, for example, for service delivery or monitoring purposes. When a registered person no longer requires that service the information can become out of date, but remain within the administrative system, sometimes for many years. Therefore, it can be difficult to produce estimates from these as we deal with people who are not registered at their usual residence and those who move.

Although the SPDs provide us with an approximate count of the population, they are subject to both missingness (under-coverage error) and incorrect inclusion (over-coverage error). Under-coverage occurs when some members of the population are inadequately represented (population not being present on administrative data, or present in the wrong location). Over-coverage error occurs when a member of the population is counted more than once at the same location (duplicate person response at the same location), more than once at a different location (duplicate person response at a different location), counted in the wrong location or is incorrectly included **(Račinskij and Hammond, 2019)**. Someone may be legitimately on an administrative source to receive a service when they were previously or temporarily resident, but still classed as incorrect inclusion for our definition of someone who is usually resident.

There are several possible mechanisms by which records may be incorrectly included in the SPD. For example, a person registering to have a National Insurance number is not necessarily a usual resident. There may also be lags in individual data sources, which may result in activity being associated with the wrong time period. It is also possible that the methods of integrating and linking sources together may introduce over-coverage and under-coverage errors. Some duplicate returns from the same location and erroneous records that can be identified are resolved during data processing. Other types of over-coverage, however, remain in the estimated population and need to be accounted for during estimation. Methods to address over- and under-coverage will be described in Section 4.

Complexity of population estimation with SPDs: the contrast with Census estimation

Census over-coverage must address duplicates and misplacement, and **reliable methods have been tested and used for these purposes** **(ONS, 2011c)**. Because of the nature of the administrative data they are based on, estimation using SPDs will involve additional types of over-coverage. The SPD list includes people who are not part of the "usually resident" population, but so far, we can only assess the net coverage through comparisons to mid-year estimates, which are affected by substantial uncertainty themselves, **and to Census 2011**. Administrative data coverage has changed substantially since 2011 and some comparisons have now been made between SPDs and Census 2021[2].

Additionally, the extent of under-coverage on SPDs is likely to be much greater than for census. Census response rates are **very high** **(ONS, 2021b)**, and this is important when making assumptions about independence of lists in estimation – the greater the coverage, the smaller the impact of violating this assumption. In some demographics and geographies, the SPD coverage is likely to be much worse, for example where migration rates are high.

[2] The ONS Census for England and Wales 2021 has been linked to the DI records for 2021 to enable a comparison of the over-coverage and under-coverage of SPD versions 3 and 4.2 relative to census, but results are not yet available. This analysis will help to inform our estimation methods and identify any special populations that are specific causes of concern for SPD over-coverage. This may include, for example, emigrants, short term migrants, and temporary visitors with temporary UK addresses.

Census also has the advantage of being followed around six to eight weeks later with the Census Coverage Survey (CCS), which enables assumptions to be made about a closed population. Some kinds of internal movers between areas are not problematic, but immigration and emigration will cause problems if they occur between the data collection for two lists being used for estimation.

In using SPDs for estimation, we must accept that the records on the SPD may have been updated a significant time before the reference date of interest at the end of June. Different sources lead to varying degrees of time difference and the SPD is tied to these. If we intend to use a continuous household survey for the population coverage survey, we must also accept that this data collection will be spread over time. It will be possible to ask respondents questions about their residence and status on a specific date in the past, but this will be subject to an unknown recall bias.

People living at the same address, as recorded on administrative data, do not necessarily form a household. We may therefore also have difficulty in practice implementing any estimator that works at the level of households.

Implementing a population coverage survey

In terms of the practical implementation of a population coverage survey within the transformed system of population statistics, it is possible that estimates or data from a population coverage survey could be fed directly into the DPM (Figure 1 Option i). This could offer some advantages, as estimation of coverage would be considered at the same time as reconciling those stocks estimates with flows. However, it is not currently the preferred option due to challenges in integrating this with the DPM's Bayesian methods. In practice, (Figure 1 Option ii) an estimation method will work by attaching two weights to each individual on the SPD: one to account for their probability of being subject to under-coverage and the other for over-coverage. The DPM could take as an input the SPD counts adjusted by these weights, or it could take the raw SPD counts and the aggregated coverage adjustment ratios.

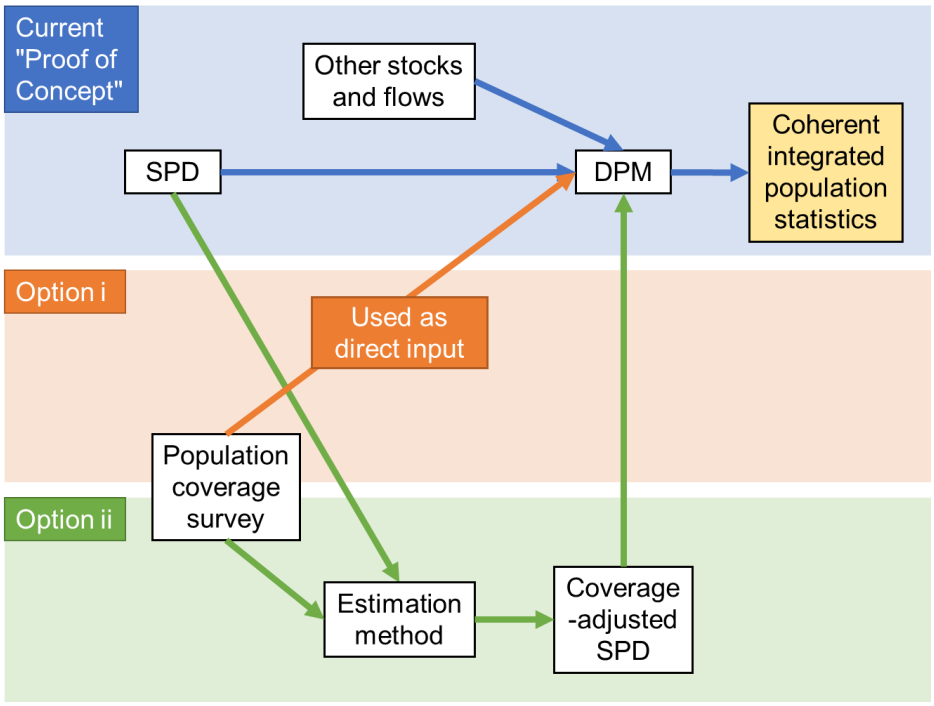


Figure 1: How data from a population coverage survey may be used to ensure DPM estimation is unbiased as administrative data coverage changes over time

In this paper, we propose possible combinations of methods and data sources that could be used to provide less biased population estimates by age, sex and LA, as required by the DPM. There is further work required to understand whether the limitations of these methods will prevent us from reaching the quality required for ongoing population statistics in the absence of a census, but we will discuss the challenges and any possible solutions that can be explored.

3. Available data sources

In Table 1 we describe the current data sources available to use in creating population estimates from SPDs. More complete information including coverage and response rates can be found in Annex 2. For use in the DPM, historical, present, and future SPDs will require adjustment. The option that is best now will be dependent on currently available data sources. This may differ from the optimal solution that would be used when new surveys and systems have been put in place.

Table 1: Available data sources

Data Source	Description	Usage in this project
Census	Census data (either unadjusted or adjusted)	Currently using 2001 in simulations, planning to use 2021
Census Coverage Survey (CCS)	Random sample of postcodes covering approximately 350,000 households stratified by Hard-to-Count (HTC)	Planning to use 2021
Labour Market Survey (LMS)	Systematic unclustered issued sample of up to 182,000 addresses per quarter, to be reduced to 142,000 with Knock-to-nudge.	Planning to use all available years
Annual Population Survey (APS) (an extension of and dependent on LFS)	Achieved annual sample of approximately 180,000 household addresses.	Planning to use if required for historical SPD estimation
Demographic Index (DI) (versions 0.4, 2.0, 2.1)	ONS composite dataset produced through linkage of PDS, CIS, HESA, ESC, WSC, ILR[3] and Births and Deaths records. Each version has different extracts of data with v0.4 also having different index ID's	Planning to use
Statistical Population Dataset (SPD) (version 2, 3, 4)	SPDv2 PR[4], CIS, HESA SC. Inclusion requires being linked on 2 sources (under 4 only PR required)	Not planning to use
	SPDv3, includes HESA, Births, PDS, ESC, WSC, BIDS/CIS. Uses DI v0.2 for 2016-19 and DI v0.4 for 2020. Inclusion rules are activity-based: individuals have interacted with one or more data sources in the 12 months prior to the mid-year reference point; Inactive relatives: They have not interacted with a data source personally but are related to and live with someone who has.	Planning to use for 2020
	SPDv4 includes sources on SPDv3 and births, HES (Hospital Episode Statistics), ECDS (Emergency Care Data Set), ILR. Currently “Presence and activity” and “Income activity” are both being worked on for inclusion rules.	Planning to use, starting with 2021

[3] PDS – Patient Demographic Service; CIS – Customer Information System; HESA – Higher Education Statistics Agency; ESC – English Schools Census; WSC – Welsh Schools Census; ILR – Individualised Learner Record

[4] PR – Patient Register

DVLA Data	Admin data of individuals on the DVLA database consisting of over 48 million driver records	Currently using totals by age-sex only
Electoral roll	It has been suggested that electoral roll could be used as a list B in some estimators. It is currently used in the construction of the DI.	Not planning to use

Should we consider any other data sources?

Should we reconsider any sources we have currently labelled as “Not planning to use”?

4. Available methods

Previous work has tested a range of methods by using simulated survey and administrative data and helped to confirm some strengths and limitations of each. There are also proposed methods that have not been tested. We summarise the methods that we are aware of in Table 2. Annex 1 describes how we are building on the previous simulations carried out to explore the properties of estimates produced by some of these methods.

Are there other methods that we should be considering?

Table 2: Estimation methods

Method	Description	Addresses	Strengths	Key assumptions	Other limitations	Bias	Survey requirements
1) Dual System Estimator (DSE)	Capture-recapture approach which observes how many records are found in two counts of the population, admin data and survey. Chapman correction estimator estimates population for given strata from the DSE, whilst correcting for small samples.	Under-coverage	Robust against non-response to the survey and household non-response It has previously been used on census adjustment, so it is well understood in terms of bias etc.	There is perfect matching between the two lists. There are no erroneous records on either list. For at least one of the two lists, non-response is homogenous within strata. Inclusion in one list is independent of inclusion in the other list. The population is closed (there is no immigration or emigration).	Several assumptions such as perfect matching and no erroneous records between admin data and survey	Violations of assumptions will result in biased estimates Any over-coverage (even if net coverage error is low) will lead to a biased result.	Survey data linked with admin data at an individual level
2) DSE Logistic Regression Under-coverage estimation for the 2021 Census of E&W (**Racinskij, 2018**)	Each individual with given characteristics has an under-coverage weight associated with them. These weights are then summed across domains of interest to estimate the population total, adjusting for under-coverage error.	Under-coverage	Robust against non-response and better performance when survey sample counts are low The logistic regression helps deal with variance	Same as DSE plus assumptions of logistic regression	Model selection can be complex.	Violations of assumptions will result in biased estimates (same as DSE)	Survey data linked with admin data at an individual level. Both the survey and population datasets should include the same key variables for modelling
3) Logistic Regression Over-coverage estimation for the 2021 Census of E&W (**Racinskij and Hammond, 2019**)	Each individual has an over-coverage probability associated with them. The over-coverage probabilities are multiplied by the under-coverage weights for each individual and summed across domains of interest to estimate the population adjusting for under-coverage and over-coverage error.	Over-coverage	Produces high quality estimates for the level of over-coverage across England and Wales. The regression approach enables key characteristics of overcount individuals to be modelled and allows for these differences across individuals to be included in the overcount probabilities.	Same as DSE plus assumptions of logistic regression	Model selection can be complex.	Violations of assumptions will result in biased estimates (same as DSE)	Survey data linked with admin data at an individual level. Survey data assumes correct location of individuals response and therefore incorrect enumerations depending on location can be determined.
Method	**Description**	**Addresses**	**Strengths**	**Key assumptions**	**Other limitations**	**Bias**	**Survey requirements**

4) Ratio estimator (**ONS, 2011a**)	Used in combination with estimates from DSE for population size estimation for a required level of geography and characteristics.		Produces population size estimates at the higher level by applying the ratio estimator from low levels of geography from the DSE to SPD counts (as an auxiliary variable) for higher levels of geography. The advantage of estimating at a national level is that people in the wrong place but still within the population will no longer appear as over-coverage or under-coverage.	Response probabilities at higher levels of geography are homogeneous to the response probabilities of those used in the DSE at lower levels.	If one uses (for example) DSE at national level the assumption of homogeneous inclusion in any survey may not hold, even within age sex groups, due to geographical variations. One could try and resolve this using some additional strata, such as LA type, but if people are in the wrong place they may be in the wrong strata. it is unclear how big an issue this may be.		(used with DSE)
5) Synthetic estimator (**Baffour et al., 2018**)	Used in combination with DSE and Ratio estimator to estimate the population size for the level smaller than the level at which we fitted the ratio in the ratio estimator.		Enables us to produce estimates at desired levels and for areas smaller than those fitted in the ratio estimator.	Response probabilities for the strata we are estimating are homogenous to those from the Ratio Estimator. This is because we use the coverage rates from the ratio estimator strata to estimate at lower levels.			(used with DSE)
6) Weighting Class **(Abbott et al., 2015)** and **(Lohr, 2021)**	Calculates a class weight for households with similar characteristics to adjust for survey non-responses The important difference between this method and DSE is that we are considering households rather than individuals and there is a need for an independent, “correct” source on which the weighting class is based.	Under-coverage of households Over-coverage within households	Robust performance against over-coverage within households whilst still dealing well with non-responding households.	All addresses are on the address frame. Addresses are correctly matched to the address frame. Response propensities are homogenous (similar) within classes. Over-coverage patterns are the same for the survey responding and non-responding households. There is no within-household non-response on the survey.	Does not deal with individual non-response within households. Assumes similar characteristic households will respond similarly and the survey captures them correctly. Can suffer from high variance (needs more research)	Violation of homogeneity of groups could lead to a positive bias Within household non-response leads to negative bias	Doesn’t require individual level linkage between survey and administrative data but does require a good household address frame and household level linkage between the survey and the administrative data. Note for administrative data we have the challenge of UPRNS not being the same as households.
Method	**Description**	**Addresses**	**Strengths**	**Key assumptions**	**Other limitations**	**Bias**	**Survey requirements**

7) Multiple system estimation (**Baffour et al., 2013**).	Similar to DSE but uses more than two lists. Use log linear models or an approach suggested by **Bishop et al., (2007)** to estimate the missing from all cells.	Under-coverage	Multiple lists should improve the results. Can deal with correlations between data sets. Pairwise dependence between lists can be modelled so this does not require an independence assumption.	Closed population of lists. Homogeneous capture probabilities for at least one of the lists. Perfect linkage between lists. No over-coverage in any of the lists. (Note: No independence assumption required here)	Requires a good understanding of the different data sources and how they may be correlated Complexity quickly grows as additional data sets are added. The assumption of no over-coverage is more likely to be violated when more lists are included. More susceptible to over-coverage due to the larger number of cells in the model.	Can model dependence bias, which reduces bias in MSE estimates compared the DSE estimates when dependence is present.	Requires multiple data sources that can be linked together.
8) Trimmed DSE	DSE cannot normally deal with over-coverage, unless the record linkage between two lists can be used to estimate the level of over-coverage. It therefore might be valuable to remove as much over-coverage as possible before using an estimator.	Removes most over-coverage so that DSE (or another method) can more successfully address under-coverage.	Assuming you can trim with better than random accuracy, trimming should improve results.		There will (almost) always still be some over-coverage. Although certain metrics have been designed to tell one when to stop trimming, it may not be clear when to stop, meaning that variance would be increased due to smaller sample size without reducing bias. It must be possible to order the list by predicted inclusion probability.	If trimming is effective, bias should be reduced compared to simple DSE.	
9) Patrick Graham Bayesian methods	Based on work by Patrick Graham in Statistics New Zealand Previous work in ONS has used a frequentist approach to the Bayesian backcalculation method.	Over-coverage and under-coverage			Assumes full survey response so artificial adjustment must be made for any form of survey non- response. High variance of estimates.		Survey data linked with administrative data at an individual level.
Method	**Description**	**Addresses**	**Strengths**	**Key assumptions**	**Other limitations**	**Bias**	**Survey requirements**
10) Fractional Counting	A model is built to predict the probability of someone	Any kind of over-	Produces fractional weights that can be	Sample data for training the model is drawn from the	It may be very difficult to collect the information	Bias could arise from any	Requires known positives and

	on administrative data being a usual resident (as opposed to being an over-coverage case). It can be trained using a dataset providing real labels of over-coverage cases through linkage to a definitive source (e.g., Census or a sample survey)	coverage for which we have training data (e.g., misplacement, emigration)	used in the same way that census over-coverage estimation works for duplicates and misplacement	same distribution as the population to which the model is applied to predict probabilities.	from over-coverage cases to fit a model for how they are different from usual residents. Variables used in modelling may not be able to fully account for difference in probabilities of being a usual resident.	problems with representativity of the training data, e.g., specific types of over-coverage case are less likely to appear	negatives to identify over-coverage cases. This could be collected through dependent sampling/interviewing from SPD
11) Latent Class Analysis	A model is built to construct hidden classes that explain the variation seen in the appearance or absence of individuals on specific sources. In some circumstances the class membership may be a useful indicator of inclusion in the population of interest.	Over-coverage	No training data is required to fit the model as it is an unsupervised method.	Non-parametric LCA has no assumptions. Categorical or Ordinal data.	Interpretation can be difficult here. There is no guarantee that the hidden classes found will correlate with presence or absence in the population of interest.	Very likely to biased in some way, as it is not fitted to a specific definition but inferred from the structure of the data.	None

5. Options

We suggest there are three high-level options to estimate the population size of England and Wales in the absence of a traditional census[5]. The options have differing requirements from the administrative population dataset used.

SPDs in their current form are most suitable for Option 1, and therefore this represents the fastest implementable solution.

> **Option 1: Register with under-coverage and over-coverage error**
> A population or address register with under-coverage and over-coverage error. This will include using both an area-based sample, where individuals in the sample are surveyed to estimate the level of under-coverage and a dependent sample, where individuals in the sample are surveyed to estimate the level of over-coverage.

It may be possible to produce an SPD version (similar to the DI) suitable for Option 2, as an iterative improvement avoiding the requirement for a costly area-based survey.

> **Option 2: Register with over-coverage and negligible under-coverage**
> Either a population or address register with negligible under-coverage, but which does include over-coverage error. Therefore, a large dependent sample can be drawn across England and Wales and individuals within the sample interviewed to estimate the level of over-coverage error.

Finally, in the future but dependent on improved administrative data quality and processes, Option 3 would be the gold standard option.

> **Option 3: Central Population Register (CPR)**
> To produce census-like estimates using a high-quality list of usual residents within England and Wales with reference to any time point. A small survey can also be used to audit the CPR.

Options 1, 2 and 3 could be delivered by using both data sources already available within the ONS, and new data collection or acquisition.

We have identified one method that would make use of data sources already available:

- Record linkage of the LMS to the SPD for under-coverage and over-coverage estimation

We have identified several methods that would require new data collection or acquisition:

- Record linkage of an SPD coverage survey to the SPD
- Record linkage of an SPD-Dependent Sample Survey to the SPD

[5] Our aim here is to focus on estimating coverage errors for individuals in private households in SPDs. We are not currently focussed on estimation for communal establishments.

- Record linkage of an administrative data source that has negligible over-coverage to the SPD to determine SPD over-coverage
- Register with negligible under-coverage
- Central Population Register

The above methods require high quality record linkage between lists. With the variables available, this linkage is unlikely to meet the same quality standards achieved by Census-CCS linkage but should involve clerical review.[6]

Considering the three high-level options we have identified for estimating the population size of England and Wales, and the above list of methods that could deliver against these options, Table 3 provides an overview of how these methods could be applied in combination for delivery of the different options. We have also identified an additional option, record linkage of a large compulsory survey and accompanying coverage survey, which we do not give further consideration to in this paper (some preliminary thoughts can be found in Annex 4).

Table 3. Methods and data sources that fulfil requirements of Options 1, 2 and 3

Option	Variant	Delivery combination
1	A	LMS
	B	LMS + Administrative data without over-coverage
	C	LMS + SPD-dependent sample
	D	Area-based SPD coverage survey + SPD-dependent sample
2		Register with negligible under-coverage + SPD-dependent sample
3		Central Population Register + LMS

In section 5.1, we provide more detail on the alternative methods for delivering against Option 1. In sections 5.2 and 5.3, we provide more detail on the methods for delivering against Options 2 and 3 respectively.

5.1 Option 1: Register with under-coverage and over-coverage error

A. LMS

Record linkage of the Labour Market Survey (LMS) to the SPD for under-coverage and over-coverage estimation (making use of currently available data sources)

The first step is to link the full SPD with the LMS under high quality record linkage requirements, which should mirror those between the Census and CCS. The linkage requirements for Census 2021 were that there should be no more than 0.1% false positive and no more than 0.25% false negative record links **(Shipsey, 2020)**. The record linkage quality requirements should scale with the quality requirements for population estimates. If quality requirements for population size estimates are high, then this will result in high quality requirements for record linkage. This will enable us

[6] It is important to note that not all these methods have been tested and therefore they are initial options to explore population size estimation with the highest level of accuracy.

to determine if an individual is counted as they are in the SPD, if they are a duplicate individual, and/or if they are counted in the wrong location.

- Benefits:
    - LMS design also allows collection of data for other purposes, therefore it is cost and time efficient.
    - The systematic sample design will allow flexibility to estimate different geographies.
    - The LMS is in development so we can suggest changes that would improve its utility for SPD estimation.
- Limitations:
    - This record linkage exercise will not be able to capture some types of over-coverage that will exist in the SPD, such as those who show signs of life, but are not usual residents.
    - Estimation methods will make strong (and maybe unrealistic) assumptions that:
        - The survey population (usual residents at a reference time point) is closed. The LMS is collected all year round and there will be movement in and out of the population throughout the year. Those who respond are asked to refer to specific quarterly dates. This may introduce recall bias.
        - The survey determines the correct enumeration of records. If the LMS asks about residence on the reference date for the SPD, there will be some recall bias.
    - 40 to 45% expected response rate. If the under-coverage of the SPD is small, the impacts of lower response rate for the survey are minimised as the unknown count for those missing from both lists will be smaller. However, a low response rate will decrease quality of estimates where dependence and over-coverage are present.
    - Using AddressBase as the sampling frame for under-coverage estimation assumes that AddressBase is perfect. In reality, we are unable to find addresses that do not exist on AddressBase and this will result in underestimating the under-coverage in the SPD.
    - It is unlikely we will meet the same linkage requirements as those in the Census. This is because of a lack of variables within the SPD to produce high quality record linkage.
    - As the LMS is an existing survey, suggested changes may not be able to be implemented for coverage estimation.
    - The systematic sample design, like proportional allocation, could be a disadvantage for estimates of smaller geographies.

Once the matching exercise is complete, individuals counted in the LMS but not the SPD within the correct estimation domain will be undercount. Individuals who are a match within the correct estimation domain will be correctly counted on the SPD. Individuals who were a match between the two lists but in a different estimation domain will be overcount individuals on the SPD.

Once the LMS and SPD are record linked, an under-coverage estimation method can be chosen, and one or more over-coverage estimation methods. Details of possible estimation methods are described in Annex 3.

As noted, the methods above using solely LMS and SPD as currently available have some severe limitations. Additional data sources would provide opportunities to use improved methods and account for more types of under- and over-coverage with increased accuracy.

B. LMS + administrative data without over coverage

Record linkage of an administrative data source that has negligible over-coverage to the SPD to be able to determine who should not be in the SPD

It may be possible in future to obtain timely administrative data that contains no over-coverage, or that can be accurately filtered to contain no over-coverage. In reality there will be some over-coverage; but if it can be minimised so that it falls within the errors required for the quality of population estimates then this would be sufficient. Investigations of trimming and filtering of SPDs have not been able to achieve this so far, but it is possible that other sources may enable it.

- Benefits:
    - Administrative data is collected and accessible, cost and time efficient.
- Limitations:
    - It will be difficult or impossible to find an administrative data source that does includes only usual residents, alongside key variables needed for high quality linkage.
    - For DSE-like methods, homogeneous capture is required for one of the lists, i.e. all members of an estimation stratum have an equal probability of being present on the list. This is unlikely to hold for administrative data sources and will introduce bias.

C. LMS + SPD-dependent sample

Record linkage of an SPD-Dependent Sample Survey to the SPD

A new SPD-Dependent Sample Survey would interview a sample of households drawn from the SPD. The SPD address list will be dependently sampled and for those addresses sampled, individuals will be interviewed to determine if they are overcount cases in the SPD. This sample would aim to capture “hard to cover” individuals, by oversampling individuals with specified characteristics who are likely to be overcount. These groups can initially be created using information from both the SPD and Census 2021 on who are most likely to be overcount individuals. Depending on the requirements for this survey and the levels of over-coverage in the population, it could either be a large dependent sample or a small dependent sample.

To make use of this data, it would be linked with high quality to the SPD. Once this record linkage exercise is complete, it will allow us to identify overcount individuals on the SPD in the sample areas. This may also enhance our understanding of the types of over-coverage within the SPD such as erroneous responses.

- Benefits:
    - Sample can be targeted to capture overcount individuals with given characteristics.
    - More accurate estimation of over-coverage, resulting in less bias in estimates.
- Limitations:
    - Dependent interviewing is not considered ethically acceptable, so we are likely to rely only on straightforward data collection about current (and potentially former) residents of a sampled SPD address only.
    - Depends on initial information about where over-coverage is most present when designing the survey, which can initially be provided by Census 2021 and later be updated over time as the survey collects information.

Methods to implement once the SPD-Dependent Sample and SPD are record linked would mirror the methods proposed for over-coverage estimation from the record linkage between the LMS and the SPD (see Annex 3). Use of the SPD-Dependent sample to estimate over-coverage would complement the existing ability of the LMS to estimate under-coverage.

D. Area-based SPD coverage survey + SPD-dependent sample

Record linkage of an area-based SPD coverage survey to the SPD

A specific SPD Coverage Survey could be designed to mirror the CCS **(ONS, 2012)**. This would be an area-based survey, which allows for addresses where individuals are not included in the SPD to be enumerated and does not rely on an address frame. It would aim to capture hard to reach individuals by using some type of “hard to cover in administrative data” indicator. This survey would be collected for a fixed length of time, as often as population size estimates are to be produced with high precision and accuracy, starting at the reference date for the SPD to maximise response but minimise movement within the population. Unlike the LMS, where the expected response rate is around 40 to 45%, we would aim to achieve a high response rate of around 90% to minimise the number of individuals missing from both lists when the lists are record linked. This may only be achieved by making the survey compulsory, using face-to-face interviewing, and increasing collection time. Depending on the requirement for this survey, it can either be a small survey of around 1% of the population to adjust for levels of SPD under-coverage comparable to census, or a larger coverage survey to adjust for larger levels of under-coverage in the population dataset.

This survey would be linked to the full SPD under high quality requirements. The main use of this survey would be to estimate under-coverage, but some types of over-coverage will also be picked up (wrong location and duplication).

- Benefits:
    - The survey can be designed to estimate the level of under-coverage for the SPD, knowledge we have from the SPD and other data sources to target harder to reach individuals.

    - This method does not depend on a sampling frame of addresses, which may itself contain under-coverage.
    - The survey can be designed to make the closed population assumption reasonable.
- Limitations:
    - Time, cost, and resource expensive.
    - Assumes closed population, homogenous response probabilities, and independence of inclusion in the two sources, which can be difficult to design and implement

Undercount and overcount individuals on the SPD will be identified in the same way as described previously for linkage of the LMS (see Annex 3). Methods to implement once the SPD Coverage Survey and SPD are record linked would mirror the methods proposed for under-coverage and over-coverage estimation from the record linkage between the LMS and the SPD (Variant A; see Annex 3).

The area-based survey and SPD-dependent sample survey suggested here follow the approach used by Italy, Israel, and other countries around the world who have transitioned to an admin-based census accompanied by coverage surveys **(Bernardini, et al., 2022a)**. The coverage survey would be designed to reduce violation of the DSE assumptions. However, this element may not represent value for money, as it is similar to using the LMS, but it may not result in a substantial improvement in coverage estimates.

5.2 Option 2. Register with over-coverage and negligible under-coverage + SPD-dependent sample

If we can use a register with negligible under-coverage, only over-coverage within the population dataset needs to be estimated and adjusted for. This approach mirrors the approach used by Israel to estimate the population size **(Zhang, 2022)**. This register can either be a population register or an address register. In England and Wales, the address register is likely to be most practical. Population data from an SPD or the DI can be joined on by address matching. A sample survey can be taken from the address register to find out if residents have different characteristics from those at the address in the SPD. The estimate of the true population could come from the ratio of those truly present to those on the SPD or a more sophisticated modelling approach.

- Benefits:
    - No under-coverage estimation is needed for population size estimation and therefore no corresponding coverage survey required
    - Time, resource and cost effective, as no survey or estimation is required for under-coverage
    - Provides an opportunity to provide feedback to admin data providers on the types of individuals on their sources who are not usually resident.
- Limitations:
    - Requirement to create a register with little or no under-coverage

- Requires a register-dependent sample to measure over-coverage
- Difficult to determine what level of under-coverage is negligible
- May still require a small area-based under-coverage survey to estimate the very small level of under-coverage in the register.
- Non-response at addresses is likely to be more likely for vacant addresses, and it is not clear how this could be adjusted for.

5.3 Option 3. High quality Central Population Register + LMS

This would provide a list of usual resident individuals within England and Wales with respect to any reference time point. This approach mirrors that used by several European countries who use this list to regularly produce census-like population statistics as outlined by **Zhang (2022)**.

- Benefits:
  - Includes all usual resident individuals in the population
  - Enables population estimates to be produced directly from the register which is resource and cost efficient
- Limitations:
  - Current SPDs are not high enough quality to be used in this way in terms of coverage and characteristics.
  - Construction of such a register would require political and public support for a universally applicable registration number, legal requirements to notify authorities when moving, and significant set-up costs and time.
  - Continuous auditing of the register would be required, but this would be less costly than larger coverage surveys.
  - May produce less accurate population estimates for low level strata as seen by Israel CPR **(Pfeffermann et al., 2019)**.

Are there any other approaches that would be suitable? Should any of these suggestions be ruled out?

6. Recommendation

At the present time, we consider Option 1 to be most suitable for population estimation for England and Wales. The current versions of SPDs have substantial over-coverage and under-coverage, therefore it is appropriate to use an estimation method for each. With the data sources currently available, Variation 1A is already feasible. The LMS could be used in this way to measure under-coverage, but with the limitation of using AddressBase and not sampling households of multiple occupancy (HMOs). Over-coverage estimation will be acceptable for internal moves between estimation domains, and other kinds of over-coverage have the potential to be estimated more accurately with changes to the LMS or using a new SPD-dependent coverage survey (Variation 1C). An area-based coverage survey could be added (Variation 1D), but the set-up and operation of such a survey would be costly,

and while this example has been used with some success in Italy, they departed from it during the pandemic and are considering other options less reliant on extensive data collection **(Bernardini, et al., 2022b)**. Variation 1B relies on an administrative source of a kind that we do not currently expect to have, but it should be brought into consideration if this changes.[7]

Of the three options we have discussed, Option 2 is recommended as the most sustainable and affordable whilst maintaining high quality, given that it does not require a substantial, ongoing area-based survey for under-coverage. Option 2 has not been thoroughly researched, however, and will require further work to confirm that it is feasible with our current data supplies with the addition of a carefully designed SPD-dependent over-coverage survey. If an address register is used as the list with negligible under-coverage, the coverage will need to be investigated, and any heterogeneity in low levels of under-coverage may be a concern. In the case of a person-level register, the DI may already be quite close to fulfilling this requirement. Results from the analysis of linkage between the DI and Census/CCS will give a useful indication of remaining under-coverage of the DI (usual residents who have not interacted with administrative systems).

Option 3 would be desirable if there were political and public will to move to using a unique registration number for all individuals, to be used when interacting with any public service. We do not expect that this to be considered in the near future for only statistical reasons.

We conclude the paper with an overview of our planned future work to explore these options further.

Are there any other options we should consider? Should we focus our future research on a subset of these options, or aim to investigate all of them with less thorough simulations?

7. Future Work

Work is already in progress on a range of simulations to reproduce and scale up previous work, adding new features to capture some of the challenges. We are initially simulating populations and administrative data using Census 2001 data from the same LAs used in previous ONS research (**North, 2022**) and described in **previous MARP papers** **(Archer et al., 2020 and Archer et al., 2021)**. It will be important to understand how estimators perform in a range of coverage scenarios for

[7] The methods we discuss here are most suitable for producing estimates for private household populations but could reasonably be extended to small communal establishments. Large communal establishment populations are enumerated by different means in census, and specific data collection to monitor coverage of these in SPDs will need to be considered separately from the coverage surveys above.

future SPDs, so we will ensure our results are robust and look further into the relationships observed so far.

We have started to explore simulations using DVLA licensing data in DSE instead of a survey as a possibility for Variation 1B. However, from the aggregate totals we are currently working with, there appears to be over-coverage in the DVLA that requires further investigation. It has been suggested we could remove over-coverage by using only recently updated records, which should be possible when we start to use record level data.

For Option 1, we will simulate a number of different SPDs (with the same over-coverage and under-coverage) to explore the effect on the variance and bias of the different estimators. We also intend to simulate trimming the SPD further before using an under-coverage estimation method. We would like to introduce dependence between presence on the SPD and response to a survey, or coverage on DVLA data, but the impact of this is well understood so it is not our highest priority.

Our current simulations outside the ONS Data Access Platform (DAP) will be repeated in DAP, where we will be able to use Census 2021 as a population pool. We will also have the computational resource to scale up simulations to estimate nationally, and to test alternatives to separate estimation of individual LAs.

Linked data are now becoming available to test some kinds of estimators in a realistic way, but also to better understand the current coverage of SPDs. Census 2021 and CCS linked to the DI is now enabling detailed analysis of SPD and DI coverage errors in 2021, which is being carried out by colleagues in SSTAR. Clerical linkage has only been used on a subsample of CCS areas (for both the CCS and Census records in those areas), so weighting will be used to account for this in analysing coverage. Special populations will be one aspect of this analysis, and in due course we will need to check that any general coverage adjustments solve the problems that may arise in specific groups.

Also using this linkage, we will be able to test an approach similar to the proposed area-based survey in Variation 1D. We intend to use the 2021 SPD version 4 with CCS to test DSE. Estimates produced in this way will be compared to final adjusted estimates from Census 2021. Care should be taken in interpreting the success or failure of such a test, as it applies to only one time point and there is not the flexibility we have in simulations to measure sampling error from multiple runs.

Similarly, we will use 2021 SPD version 4.2 with other designs of simulated survey (Variations 1A, C and D), by drawing repeated samples from Census 2021. Simulations of surveys alongside a real SPD will provide sampling errors, and bias can be calculated by comparison to Census 2021 estimates. Currently, this work would be limited by lower recall (92%) automatic linkage between the full Census and DI, but higher quality linkage is in progress and due to be completed in early 2023.

To explore the potential of Option 2, our simulations and estimation methods will be extended to consider the scenario of minimal under-coverage. In this context, we can explore various possibilities for implementation of a dependent sample using a population or address-based register.

It has been suggested that longitudinal survey data collection may improve our ability to estimate some aspects of coverage, perhaps by giving more insight into churn. Currently, we are not sure what value this would add in addition to asking individuals about their location of residence at previous points in time. Following up the same individuals again would use resource on collecting data on those for whom we already have some data, whereas drawing fresh sample will reduce correlation between consecutive time points and provide a greater effective sample size when pooling of sample over time.

We intend to collaborate more closely with other NSIs using or planning similar approaches, for example, Italy, Israel and Latvia, to understand how they have overcome specific challenges we face.

Is there any other research that we should be considering? Should any of these plans be higher or lower priority?

References

Abbott, O., Castaldo, A., Racinskij, V., Ross, H., Smith, P. and Brown, J., (2015). 'Developing a weighting-class approach for the 2021 Census.'

Archer, R., North. R., Metcalfe, A., (2020). 'Estimating population size without a census'. Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP129-Estimating-population-size-without-a-census.docx

Archer, R., North. R., Metcalfe, A., (2020). 'Estimating population size without a census: Appendix results tables.' Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP130-Estimating-population-size-without-a-census-results-supplement.docx

Archer, R., North. R., Metcalfe, A., (2021). 'Population estimation without a Census: update'. Available at: EAP151 - Population estimation without a Census: update (statisticsauthority.gov.uk)

Baffour, B, Brown, J. and Smith, P. (2013) An investigation of triple system estimators in censuses. *Statistical Journal of the International Association for Official Statistics*, **29**, 53-68.

Baffour B., Silva D., Veiga A., Sexton C., and Brown J. J. (2018), Small Area Estimation Strategy for the 2011Census in England and Wales. *Statistical Journal of the International Association for Official Statistics*, **34**, 395-407.

Bernardini, A., Brown, J., Chipperfield, J., Bycroft, C., Chieppa, A., Cibella, N., Dunnet, G., Hawkes, M.F., Hleihel, A., Law, E.C.,Ward, D., and Zhang, L.C., (2022a) Evolution of the person census and the estimation of population counts in New Zealand, United Kingdom, Italy and Israel. *Statistical Journal of the IAOS*, (Preprint), pp.1-17.

Bernardini, A., Chieppa, A., Cibella, N., Gallo, G., Solari, F. and Zindato, D., (2022b) Evolution of the Italian Permanent Population Census. Lessons learnt from the first cycle and the design of the Permanent Census beyond 2021.

Bishop, Y.M., Fienberg, S.E. and Holland, P.W., (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Haldane, J.B.S (1945) On a method of estimating frequencies. Biometrika, Vol 33, No. 3, pp. 222-225

Hammond, C. and Naprta, M. (2021). The Proposed Duplication Calibration Method for the 2021 Census of England and Wales

Lohr, S.L., 2021. *Sampling: design and analysis*. Chapman and Hall/CRC.

North, R., (2022). 'Coverage estimation without a census: Simulation study for 2011'. [*internal paper available on request*]

Office for National Statistics (2011a) 2011 Census Coverage assessment: CCS sample sizes and Estimation Areas for Local Authorities. Census Advisory Group Paper AG (10) 20. Available from https://www.ons.gov.uk/ons/guide-

method/census/2011/the-2011-census/census-consultations/uag/census-advisory-groups/statistical-development/census-coverage-assessment---ccs-sample-sizes.pdf

Office for National Statistics 2011b. (2011). 'Beyond 2011: ONS Response to recommendations from the Independent Review of Methodology'. Available from: https://webarchive.nationalarchives.gov.uk/ukgwa/20160105160709/http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011--ons-response-to-recommendations-from-the-independent-review-of-methodology.pdf

Office for National Statistics 2011c. (2011). 'Measuring and adjusting for coverage patterns in the admin-based population estimates, England and Wales: 2011'. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/measuringandadjustingforcoveragepatternsintheadminbasedpopulationestimatesenglandandwales/2011

Office for National Statistics (2012). '2011 Census Coverage Survey (2011 Census Evaluation Report)'. Office for National Statistics. Available from http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/evaluation---census-coverage-survey.pdf

Office for National Statistics. (2019). 'Integrated Population and Characteristics Survey (IPACS)'. Available from: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP119-Integrated-Population-and-Characteristics-Survey-IPACS.docx

Office for National Statistics 2020a. (2020). 'Admin-based population estimates and statistical uncertainty: July 2020'. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/adminbasedpopulationestimatesandstatisticaluncertainty/july2020

Office for National Statistics 2020b. (2020). 'Labour Market Characteristics report'. Available at: Labour Market Survey: characteristics report - Office for National Statistics (ons.gov.uk).

Office for National Statistics 2021a. (2021). 'Admin-based population and migration estimates: research update'. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/adminbasedpopulationandmigrationestimates/researchupdate#:~:text=Admin%2Dbased%20population%20estimates%20time%2Dseries%20analysis,-We%20have%20two&text=Each%20method%20uses%20a%20different,population%20dataset%20(SPD)%20v2.

Office for National Statistics 2021b. (2021). '97 per cent of households respond to Census 2021'. Available from: https://census.gov.uk/news/97-per-cent-of-households-respond-to-census-2021.

Office for National Statistics 2021c. (2021). 'Electoral Statistics, UK: December 2021'. Available from: Electoral statistics, UK - Office for National Statistics (ons.gov.uk).

Office for National Statistics. (2022). 'ANNUAL POPULATION SURVEY/LOCAL AREA DATABASE'. Labour Force Survey User Guide, Volume 6. Available from: https://www.ons.gov.uk/file?uri=/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance/volume62022.pdf.

Pfeffermann, D., Ben-Hur, D., and Blum, O. (2019) Planning the next census for Israel. Stat. Trans., 20, 7–19.

Račinskij, V. (2018) Coverage Estimation Strategy for the 2021 Census of England and Wales. Report presented at the Census External Assurance Panel on 16 October, 2018. 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

Račinskij, V. & Hammond, C. (2019) Overcoverage estimation strategy for the 2021 Census of England & Wales, Office for National Statistics. Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP112-Over-coverage-estimation-strategy-for-the-2021-Census-of-England-and-Wales.docx

Shipsey, R., (2022). 2021 Census to CCS Matching (Design Authority Board). [*internal paper*].

Zhang, L.C., 2022. Complementarities of Survey and Population Registers. *Wiley StatsRef: Statistics Reference Online*, pp.1-5.

Annex 1: Simulation studies

The simulations in this project serve three main purposes. Firstly, simulations allow us to improve our understanding of the estimators in a controlled environment. Secondly, the simulations have allowed us to make significant progress whilst we wait for linkage of real data to be completed. Finally, simulations have increased malleability when compared to using the real data. This means we can model different scenarios that can encompass what the future of administrative data, or surveys may look like. We can also use simulations to compare the properties of estimators using different versions of the SPDs available now, and to suggest what changes to SPDs could improve them.

Our current work aims to produce a body of reusable code that can be adapted to simulate a range of scenarios of survey and administrative data types. We are reproducing and building on work done by others outside the ONS data access platform (DAP) in R, but now working in Python to facilitate a more straightforward scale-up in DAP. The computational power available in DAP will enable us to compare estimation at the national level to estimation of LAs independently and later aggregation. Of the options listed in the Methods Table above, we have focused on DSE, DSE logistic and weighting class estimators, as these have shown promise in previous ONS work. Previous work (**North, 2022**) has shown that the increased complexity and more stringent assumptions, such as perfect survey response rate, of models such as back calculation or any Bayesian methods do not yield advantages over the better established methods

Our work so far has simulated individual LAs. For our current simulations, we use a population pool of Census 2001 to draw households/individuals from to construct the simulated “true” population and a simulated SPD. This is in line with previous ONS work. We now have access to a cut of Census 2021 within DAP, which allows us to use an up to date population pool with coverage of all of England and Wales.

The outputs of these simulations are still a work in progress and full quality assurance of our implementation is yet to be completed. Therefore, anything from the simulations presented here should be seen as an example of the kind of tests and output we intend to produce rather than an indication of the amount of bias or variance these methods are expected to produce.

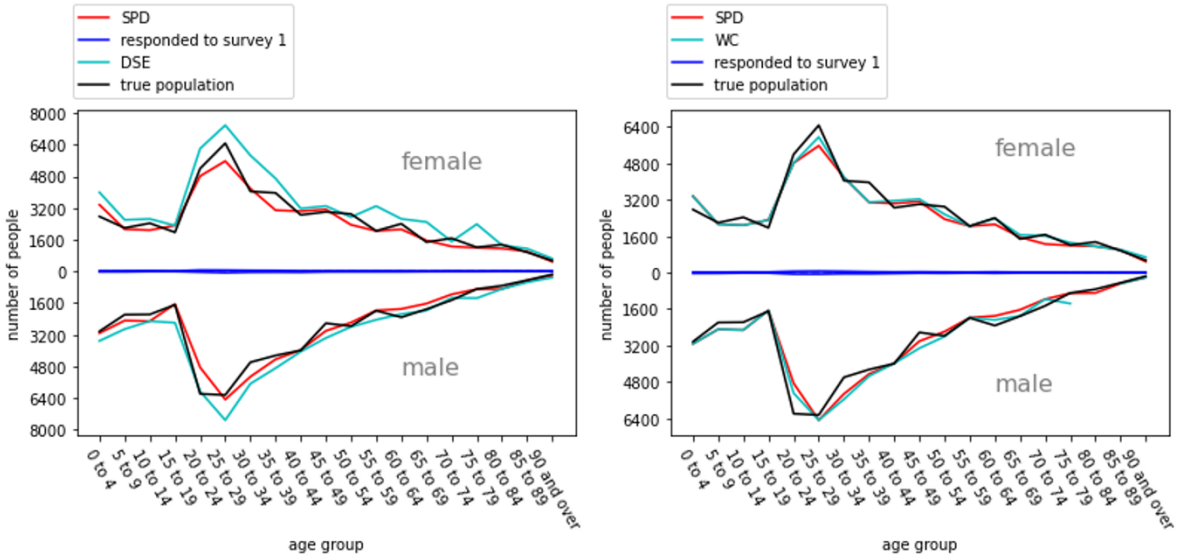


Figure A1a and b: These plots show a single realisation for the simulated surveys in one LA, showing the number of people in the true population, SPD and survey, as well as the number of people from the estimation methods, DSE on the left (Fig A2a) and weighting class on the right (Fig A2b). The plots should be seen as an illustration of how the two estimators work, rather than a comment on their properties. Here under-coverage, over-coverage, survey size and survey response rate are fixed at 15%, 15%, 1% and 60% respectively.

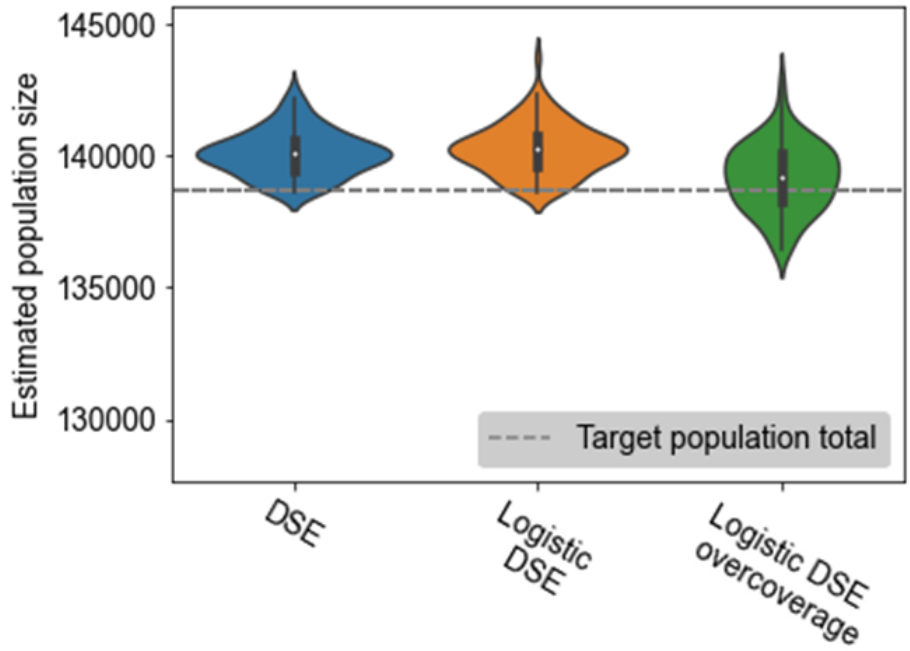


Figure A3: Violin plot of population size distributions for DSE, logistic DSE and logistic DSE with over-coverage adjustment. Simulation parameters are held constant with the following values: under-coverage = 1%, over-coverage = 1%, survey size = 0.5%, survey response rate = 50%

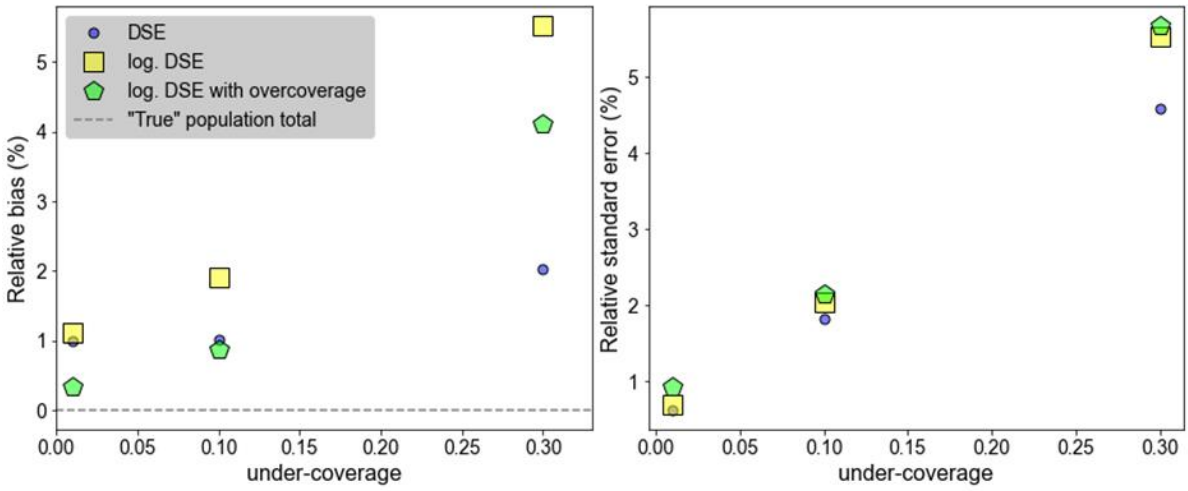


Figure A4a and A4b: These plots show how relative bias (left Fig A4a) and relative standard error (right Fig A4b) vary with under-coverage for DSE, logistic DSE, and logistic DSE with over-coverage adjustment. Here over-coverage, survey size and survey response rate are fixed at 10%, 0.5% and 50% respectively.

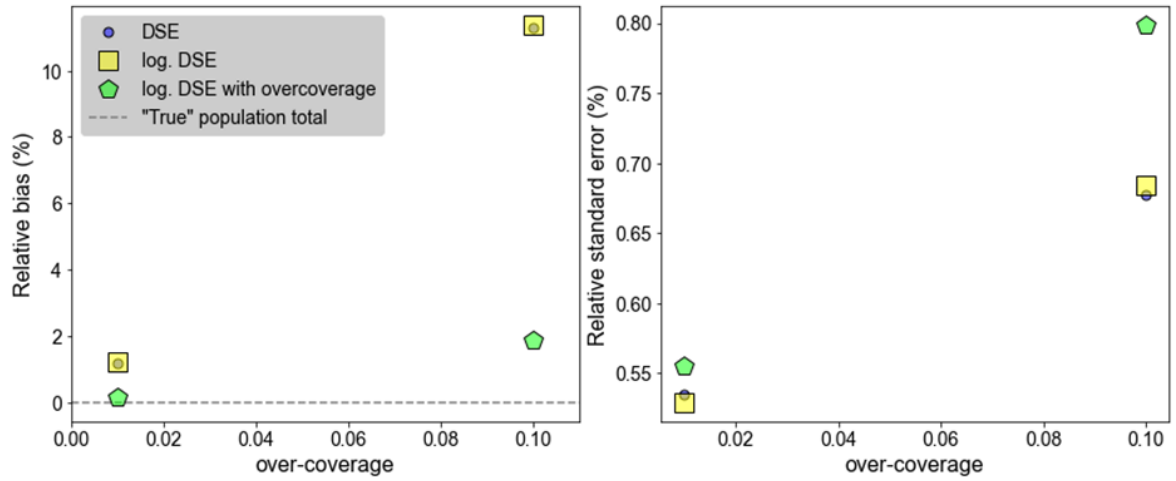


Fig A5a and A5b: These plots show how the relative bias (left fig A5a) and the coefficient of variation (right fig A5b) vary with over-coverage for DSE, logistic DSE, and logistic DSE with over-coverage adjustment. Here under-coverage, survey size and survey response rate fixed to 10%, 5% and 50% respectively.

Our preliminary results highlight the necessity of further testing. The method that appears “best” really depends on the SPD coverage and survey design. Figs A3 and A4 suggest that for some regions of the parameter space weighting class is better than DSE but for others DSE is better than weighting class (in terms of bias and variance). Again, it is important to reiterate, these should not be taken as results, but as an indication of the directions of our research.

Should we expect weighting class to have higher variance than DSE, as has been seen in previous work, but not consistently?

Are there effects or parameters we should be exploring in simulations that have not been mentioned?

Implementation

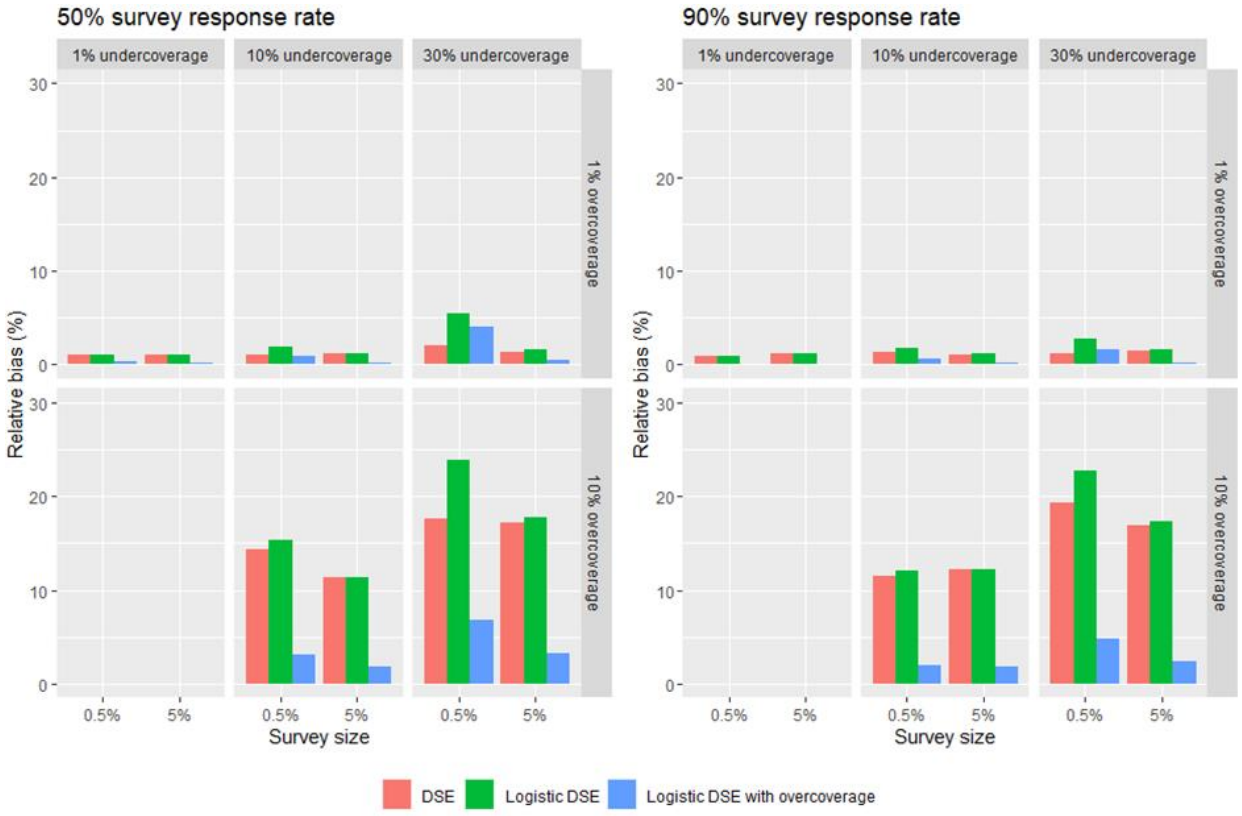
The simulations are set up so that several different parameters, such as over-coverage and under-coverage and within household nonresponse, can be dependent on age and sex. However, due to limitations on time and access to data these are currently set to a flat response with some set differences between age and sex groups determined by adding uniformly distributed random noise.

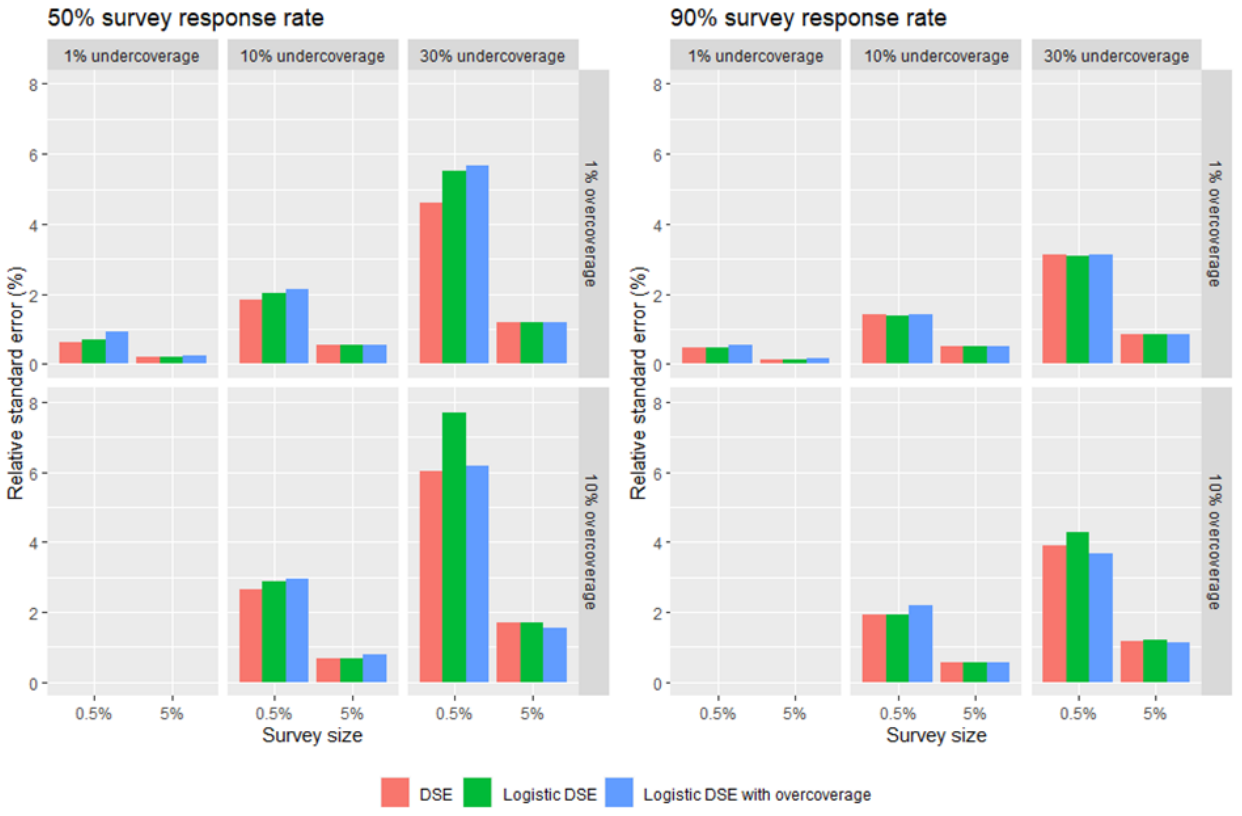
The simulations currently work by taking the population pool (a version of Census 2001) and then placing a number of these people in what we call the “true population” and then (independently) placing a number of people in what we call the SPD. The people who are in the “true population” but not the SPD are regarded as under-coverage, the people who are not in “true population” but are in the SPD are over-coverage.

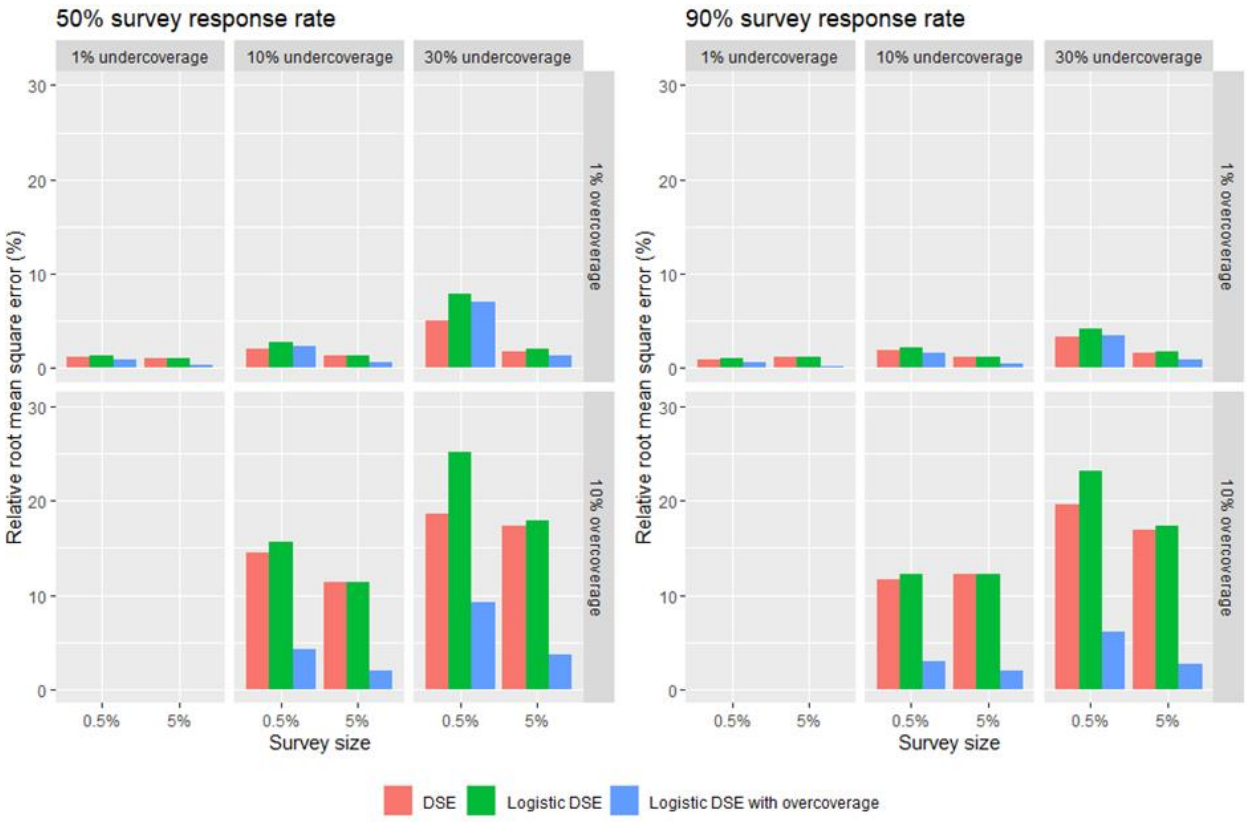
The code then simulates 100 surveys with a given survey size and response rate. We can use the SPD and the surveys to estimate the “true population”. We can then look at the variance, bias and other features of the estimates.

We have run these simulations for a number of different values for the parameters survey size, survey response rate, over-coverage, and under-coverage. The plots showing all of the bias and variance for all these combinations can be seen below with the accompanying table.

Again, it is important to reiterate, these should not be taken as results, but as an indication of the directions of our research.


50% survey response rate
90% survey response rate
1% undercoverage
10% undercoverage
30% undercoverage
1% overcoverage
10% overcoverage
Relative bias (%)
Survey size
0.5%
5%
DSE
Logistic DSE
Logistic DSE with overcoverage


50% survey response rate
90% survey response rate
1% undercoverage
10% undercoverage
30% undercoverage
1% overcoverage
10% overcoverage
Relative standard error (%)
Survey size
0.5%
5%
DSE
Logistic DSE
Logistic DSE with overcoverage


50% survey response rate
90% survey response rate
1% undercoverage
10% undercoverage
30% undercoverage
1% overcoverage
10% overcoverage
Relative root mean square error (%)
0
10
20
30
0.5%
5%
Survey size
DSE
Logistic DSE
Logistic DSE with overcoverage

Annex 2

Table A1. Additional information about currently available data sources

Data Source	Description	Coverage	Response rate (if applicable)	Years Covered	Usage in this project	Link
Census	Census data (either unadjusted or adjusted)	(2021) AddressBase	97% (2021)	..., 2001, 2011, 2021	Currently using 2001 in simulations, planning to use 2021	
Census Coverage Survey (CCS)	Random sample of postcodes covering approximately 350,000 households stratified by Hard-to-Count (HTC)	Excludes large Communal Establishments (CEs) (defined as 50+ bed spaces for 2021)	61% (2021)	..., 2001, 2011, 2021	Planning to use 2021	
Labour Market Survey (LMS)	Systematic unclustered issued sample of up to 182,000 addresses per quarter, to be reduced to 142,000 with Knock-to-nudge.	Excludes CEs and Homes with Multiple Occupants (HMOs)	37% (2021, phone/online) 40-45% expected with Knock-to-nudge	2020 onwards	Planning to use all available years	**link** (ONS, 2020b)
Annual Population Survey (APS) (an extension of and dependent on LFS)	Achieved annual sample of approximately 180,000 household addresses.	Excludes most CEs; includes HMOs		All up to present. To be decom-missioned in 2023	Planning to use if required for historical SPD estimation	**link** (ONS, 2022)

Demographic Index (DI) (versions 0.4, 2.0, 2.1)	ONS composite dataset produced through linkage of PDS, CIS, HESA, ESC, WSC, ILR[8] and Births and Deaths records. Each version has different extracts of data with v0.4 also having different index ID's	Over-coverage caused by missed links and under-coverage due to false links. Excludes individuals who have never been present on any of the core administrative data sources.	2011-2021 HESA (2016-2020) CIS (2011-2020) ESC (2016-2021) WSC (2016-2021) PDS (2016-2021	Planning to use
Statistical Population Dataset (SPD) (version 2, 3, 4)	SPDv2 PR[9], CIS, HESA SC. Inclusion requires being linked on 2 sources (under 4 only PR required)	2016-2020 up to around 10% net over-coverage (particularly working age males) and up to 10% net under-coverage (depending on age), compared to MYE	2016-2020	Not planning to use
	SPPv3, includes HESA, Births, PDS, ESC, WSC, BIDS/CIS. Uses DI v0.2 for 2016-19 and DI v0.4 for 2020. Inclusion rules are activity-based: individuals have interacted with one or more data sources in the 12 months prior to the mid-year reference point; Inactive relatives are included: they	Uses DI v0.2 for 2016-19 and DI v0.4 for 2020. V3 generally underestimates the population, (net under-coverage) but it does still have over-coverage. It will be linked to the 2021 Census/CCS to compare coverage.	2016-19 (DI v0.2), 2020 (DI v0.4)	Planning to use for 2020

[8] PDS – Patient Demographic Service; CIS – Customer Information System; HESA – Higher Education Statistics Agency; ESC – English Schools Census; WSC – Welsh Schools Census; ILR – Individualised Learner Record
[9] PR – Patient Register

	have not interacted with a data source personally but are related to and live with someone who has.				
	SPDv4 includes sources on SPDv3 plus births, HES (Hospital Episode Statistics), ECDS (Emergency Care Data Set), ILR. Currently “Presence and activity” and “Income activity” are both being worked on for inclusion rules.	SPD v4 is still a work in progress.	2016-2021	Planning to use, starting with 2021	
DVLA Data	Admin data of individuals on the DVLA database consisting of over 48 million driver records	Several issues including no unique identifier with SPD, no proof of address required and no under 16's included. All of which can cause problems and be an additional source of over-coverage and under-coverage	All up to present	Currently using totals by age-sex only	
Electoral roll	It has been suggested that electoral roll could be used as a list B in some estimators. It is currently used in the construction of the DI.	Will have bias related to those who usually vote. No under 17s. 70-75% coverage.	All up to present	Not planning to use	**link** (ONS, 2021c)

Annex 3: Estimation methods

For under-coverage estimation:

- Dual System Estimator (DSE), Ratio Estimator and Local Synthetic Estimator
    - This is the desired approach when the sample is small for the specified strata and we want to minimise the violation of the assumption of homogenous response probabilities. Steps taken:
        - Estimate the population size for the specified strata originally from the DSE.
        - Sum the DSE estimates for these strata for desired demographics and apply the ratio between the estimate and the SPD count for these summed areas to the SPD counts for higher levels of geography and characteristics.
        - For the Local Synthetic Estimator, apply the ratio of the DSE estimates and SPD counts (e.g. those at Estimation Area level with given characteristics in the 2011 Census) to a lower level SPD count, to estimate at a lower level of geography (e.g. Local Authority level with given characteristics). This approach was used in the 2011 Census of England and Wales to produce high quality Census estimates.
- Mixed effects Logistic Regression or Fixed effects Logistic Regression
    - This is the desired approach when variables that are significant in the logistic regression model are present in both lists (**Racinskij, 2018**). This is to estimate for SPD under-coverage error, where the variables used for modelling and scoring will be collected from the survey. The modelled population will be those who were in the survey after the record linkage between the survey and SPD is complete.
    - However, this approach depends on both the SPD and accompanying survey including the same key variables for coverage estimation.
- DSE only
    - This is not an appropriate method because it assumes homogenous response probabilities and that the sample size is large enough for the specified strata. Population sizes are estimated directly from the DSE under the assumptions listed in the Methods Table. The strata would have to be large enough to directly estimate here and we would assume the coverage patterns of the individuals in the strata would be homogenous.
- National level DSE and Local Synthetic Estimator
    - This is not an appropriate method, as although we can ignore wrong location overcount within the SPD, we assume homogenous response probabilities nationally within specified strata which is likely to be violated here. Steps taken:
        - Estimate the population sizes directly from the DSE under certain assumptions, nationally for certain characteristics.
        - Apply the ratio between the DSE estimate and SPD count for the strata and apply this ratio to the SPD counts for the domain of interest we want to estimate the population at.

- Weighting Class estimator (also addresses some over-coverage)
    - This estimator has not been widely used or thoroughly tested. If over-coverage is not removed effectively by other methods, it may be preferable to DSE.

For over-coverage estimation:

- Mixed effects Logistic Regression or Fixed effects Logistic Regression
    - Model an outcome variable that is 0 if a record is either a duplicate or a detectable overcount, and 1 otherwise. Independent variables can be selected from those available on the SPD.
    - The model is fitted to those who are in both the SPD and the survey, and applied to those who are on the SPD.
- Group-wise overcount propensities
    - Combine duplication and wrong location overcount and calculate propensities for groups with prespecified characteristics.
- Weighting Class estimator – only accounts for over-coverage within households.
- Fractional counting could be used for any other kinds of over-coverage for which sufficient data is collected. We think that for some types of over-coverage, it may be possible to get responses that can be used to build models similar to those used for misplacement and duplicates. This will give an opportunity to make predictions of probabilities of being over-coverage to apply to the rest of the SPD.

Annex 4

An alternative to Options 1-3: Record Linkage of a large compulsory survey and accompanying coverage survey

To overcome the problem of erroneous inclusion in SPDs, an effective alternative would be to run a large compulsory survey (LCS). This could be designed to represent the population of England and Wales, using a sample of around 5% of the total households within England and Wales. An independent coverage survey – a subsample of the LCS – would mirror the traditional census coverage survey for England and Wales, but around 20% of the size of the LCS (1% of all households). Important variables will be included in the design and collection of the LCS and coverage survey to ensure high quality linkage, the ability to extend the models to include these variables where significant and to publish estimates within domains of interest. The design of the current LMS could be suitable to fulfil the requirements of the independent coverage survey. However, data collection is continuous throughout the year and therefore does not offer the advantage of minimising movement between the two captures.

Record link the Large Compulsory Survey (LCS) and coverage survey through high quality linkage, as key variables will be included. This will allow us to identify under-coverage and over-coverage cases within the LCS using the coverage survey.

Estimate the population size for the LCS by directly estimating using a suitable DSE. Apply the ratio between the DSE estimates and SPD counts to the domain of interest for the SPD counts.

- Benefits:
    - Include key variables for linkage, estimation, and outputs
    - Similar and familiar methods used for the decennial Census of England and Wales
    - Apply ratio estimator to the SPD counts, which will mean not estimating under-coverage and over-coverage directly for the SPD
- Limitations:
    - Very high cost to implement a large compulsory survey and accompanying coverage survey, therefore not realistic to run annually.
    - Current legislation only allows a compulsory survey/census to be carried out every 5 years across England and Wales.
    - Issues around public acceptability.