**Admin Data Census Population Coverage Survey:**

**Design considerations and test plans**

**Intro**

The Office for National Statistics (ONS) is working towards transforming the Population Statistics System by 2022. Critical to achieving this is the successful development of methods for producing annual estimates of population size. Since 2013, the ONS has demonstrated potential in the use of administrative data, publishing a series of research outputs from admin data to compare with official population estimates[[1]](#footnote-2). While we continue to refine this method with the acquisition of new data sources, there is the potential need to develop a Population Coverage Survey (PCS) to assess and adjust for coverage errors on the administrative records which would otherwise result in biased population estimates.

In this paper, we set out the design considerations underpinning a PCS to support administrative data, and our plans to integrate with wider social survey transformation, most notably the ONS Labour Force Survey.

**Background to admin data**

The ONS has developed a Statistical Population Dataset (SPD) to support the development of administrative data population estimates. This is constructed by linking 4 sources together that, when combined, are expected to have broad coverage of the usual resident population. These are:

* The DWP Customer Information System (CIS) – a list of all NINo registrations
* The NHS Patient Register (PR) – a snapshot of current GP registrations
* Higher Education Statistics Agency (HESA) – students registered in the year for higher education courses
* The English and Welsh School Census (SC) – pupils registered in state schools

These datasets are characterised by both undercoverage and overcoverage of population. For example, migrants that have recently arrived in UK may not register with a GP in the first few months of their arrival. Similarly, UK residents that have not used GP services for longer periods may have been deregistered from the PR. Situations such as these result in undercoverage of the population, and an adjustment is required to account for them in population estimates.

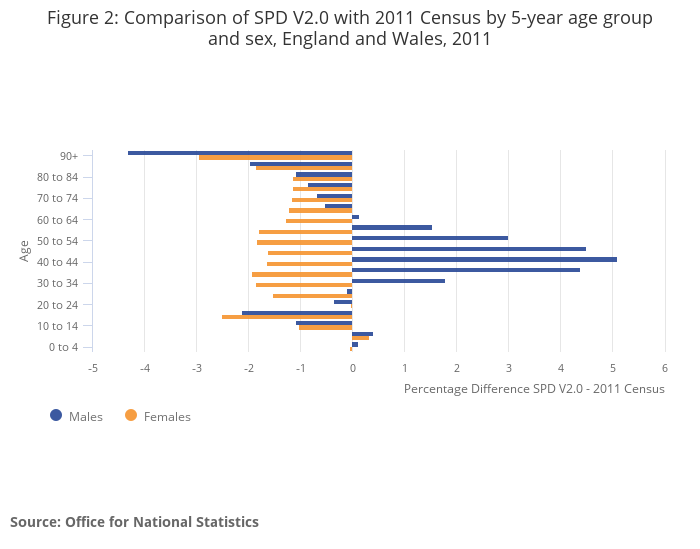
Methods to account for under-coverage are already well established in traditional censuses, for example the Dual System Estimation (DSE) approach that was used for the 2001 and 2011 Census of England and Wales[[2]](#footnote-3).

By linking multiple administrative sources to construct the SPD the level of population under-coverage is reduced. Current rules for inclusion on the SPD is for individuals to appear on at least two of the four sources listed above. We still anticipate that there is some residual undercoverage on the SPD, and depending on the availability of additional data sources that can account for persons missing on the SPD, it is likely that a similar DSE adjustment may be needed to account for undercoverage.

Overcoverage presents more of a challenge in estimating population size, as there are few established methods for adjusting for records that are erroneously captured in administrative datasets. While the 2011 Census methodology had processes in place to adjust for duplications and individuals that were captured in the wrong location, erroneous records (i.e. records that should not appear in the Census file at all) were largely expected to be negligible in number, and no specific adjustments were made. Erroneous records will be more prominent in administrative datasets, largely as their removal depends on individuals deregistering from services or notifying governments departments of departure from UK.

Figure 1 below shows the extent to which erroneous records are contributing to overcoverage on the SPD. Despite using the ‘2 of 4’ rule that requires individuals to appear on at least two of the datasets, estimates for working age males are notably higher than the Census estimates in 2011.

Figure 1: SPD 2011 Population estimates compared with 2011 Census



For the majority of age-sex groups, the 2011 SPD underestimates the population in comparison to 2011 Census estimates. However, it is important to recognise that each age-sex group will have both undercoverage and overcoverage at the record level, and the SPD estimates include the net effect of coverage errors that are cancelled out between missing and erroneous inclusions.

**Designing a coverage survey for administrative data**

The 2011 Census Coverage Survey (CCS) was designed primarily to adjust for Census non-response and achieve quality standards for the precision of population estimates at local authority level.[[3]](#footnote-4) Below is a summary of some of the key design considerations and why they may need to be different for a PCS intended to support administrative data.

*Overcoverage*: As described above, adjustment of erroneous records was not required for the 2011 Census but will need to be designed for in admin-based population estimates. This may require a separate survey component specifically designed to measure overcoverage.

*Response rates:* While the voluntary CCS achieved response rates of between 85-95%, this is largely attributed to its close association and proximity with the mandatory Census (collected within 6-8 weeks). Response rates to a voluntary PCS are expected to be lower, increasing the risk of dependence bias.

*Reference date:* The extraction of administrative data sources used to construct the SPD can be aligned to occur on a fixed reference date. However, the information itself will not be up to date for all individuals (for example, recent address moves). The PCS will need to be designed in such a way that it can account for lags on the administrative records.

*Population definitions:* Countries already transitioned to register based censuses are increasingly moving away from the ‘usual residence’ definition. ‘Resident in year’ definitions, for example, are in some instances more suitable for administrative based population estimates, and there may be benefit in designing the PCS to measure alternative definitions of population.

*Address Frame:* An ongoing address frame will be needed to support the PCS. The Ordnance Survey AddressBase product is expected to form the basis of the address frame, however the accuracy of dwelling coverage will need to be measured and potentially adjusted for in the population estimates.

*Mixed Mode Collection:* The CCS has typically relied on doorstep interviewing as an alternative to self-completion of Census forms to ensure independence in the collection of data. Since there is operational separation in the collection of administrative records, all modes of collection can be used to maximise response rates to the PCS. However, mode effects may also need to be accounted for in the estimation process.

*Questionnaire content:* Similar to the CCS, the PCS will achieve higher response rates if it is based on a short doorstep interview (10 minutes per household). However, due to the ongoing costs of running a field exercise of this scale, we are exploring the potential of integrating this survey with the ONS Labour Force Survey.

*Sample Design:* Current assumptions are that similar levels of precision will be targeted for local authority population estimates as achieved with the 2011 Census. We therefore anticipate an overall sample size of 350,000 households, with sample allocation spread across all 348 local authorities. An admin data equivalent to the Census Hard to Count (HtC) Index[[4]](#footnote-5) will be needed to reflect the variation of population subgroups missing or erroneously captured on administrative data. The sample could potentially be designed for rolling collection throughout the year, rather than collecting data over a short annual period in the weeks immediately following a fixed reference date.

*Estimation Framework:* There is potential to draw upon similar estimation frameworks used for 2011 Census (DSE / ratio estimation). However, depending on the suitability of alternative sources for detecting and removing overcoverage, it is possible that an additional over-coverage adjustment will need to be included in the DSE framework. Other alternatives include model-based approaches, which estimate over-coverage parameters (e.g. New Zealand approach)

**Integration with ONS Labour Market Survey**

Due to the complexity of design features highlighted above, the cost of running the PCS and associated field force requirements has not yet been established. It is certainly the case that an ongoing PCS, whether collected annually or continuously, would have significantly higher costs and resource implications than the ten-yearly CCS. Where possible, integration with other ONS Social Surveys is desirable to deliver value for money, provided key requirements for the design can be met.

The Labour Force Survey (LFS) is one of ONS’ largest social survey collections. Based on a wave model where respondents are invited to take part over 5 quarterly face to face collections[[5]](#footnote-6), the survey covers core labour market statistics, as well as overlapping content with Census topics. ONS are in the process of redesigning the first wave of this survey for online self-completion, with non-responders followed up with face to face interviewing. This redesigned first wave that is being tested is known as the Labour Market Survey (LMS), and content has been reduced to core labour market questions with approximately 15 minutes completion time per household. Incentivisation is also expected to increase overall response rates, and a proposed LMS sample size of 350,000 households per annum has the potential to serve as the basis of a ‘Master Wave’ which can be used as a sample frame for other ONS surveys.

A number of design features in the proposed ‘Master Wave’ present opportunities for integration with the PCS.

* Similar overall sample size targeted per annum (300k-350k households)
* Mixed mode collection (online self-completion, face to face interview, telephone)
* Full enumeration of all household residents
* Planned to be operational by 2020

A series of test plans to evaluate the viability of integrating the PCS with the LMS ‘Master Wave’ have already commenced. In Autumn 2017, response rates were compared from a split sample of 4,000 households invited to take part in a mixed mode survey. Half of the sample were asked questions that would be typically be asked in a separate coverage survey. The remaining 2,000 households were asked coverage survey questions as well as core labour market questions covered in the LMS. The results of this test are currently being analysed, however early indications are that the response rates and the quality of response is similar across the two samples.

The more significant challenge of integrating the PCS with the LMS is the development of a compatible sample design. The wave design characterising the current LFS model would continue under the LMS redesign, therefore the PCS would need to be collected on a continuous basis throughout the year, as opposed to a fixed annual collection period. This may present challenges in producing population estimates for a specific reference date, for example the current mid-year (June 30th).

More critical is the systematic sampling approach used for the LFS that is expected to continue under the LMS redesign. The LFS approach of unclustered sampling of addresses has implications for coverage surveys that are typically designed around area-based sampling. Area-based sampling is the process of undertaking full enumeration of all households in a designated area (e.g. selected postcodes). This approach ensures that all residential dwellings within the sampled areas are accounted for to make an independent adjustment for dwellings missing on the address frame.

Stratification requirements for the PCS also need to be taken into consideration. Whereas previous CCS stratification has been area based (i.e. a hard to count index at postcode level), there is potential to develop a stratification index for admin data at a more granular level. One of the advantages of having administrative records linked to the address frame, is in predicting the propensity for coverage errors at each address. There is potential to use the household composition algorithm we have developed using administrative data, as well as other socio-economic variables we have available at record level to construct a stratification index for addresses. One of the options we are considering is the use of systematic sampling within PCS strata as a means to ensure the variation of population coverage error is captured in the sample.

**Additional requirements to support an integrated PCS / LMS**

**Property Listings**

While there is potential to stratify the PCS sample at the address level to measure coverage errors, there is still the need to account for dwellings that have not been captured on the address frame. For the CCS in 2011, the first 2-3 days of field work in the sampled postcode areas was used to ensure all residential dwellings were identified for enumeration in the area. For a PCS that is integrated with the LMS design, the unclustered sampling of addresses will not lend itself to measuring coverage of the address frame that has been used to draw the sample.

We are still in the early stages of evaluating the quality of AddressBase as the basis of an address frame for ONS social surveys. Depending on the results of our field testing in Autumn 2018, we envisage there being a supplementary property listing exercise to support the PCS. This will be separate from the PCS/LMS sample and will have the sole purpose of independently collecting information on residential dwellings in selected areas. Current plans are for field workers undertaking the property listing exercise to do so ‘from scratch’, i.e. without a starting list of properties in the area. This ensures an independent view on the quality of the address frame. Once data has been collected in the field, the addresses identified in the selected areas will be compared to AddressBase. This will enable the measurement of coverage errors on the frame and any necessary adjustments to household and population estimates. Current plans are to modernise the collection process using electronic devices with GPS mapping applications. Since AddressBase includes GPS co-ordinates for all Unique Property Numbers (UPRNs), there is potential to have more flexibility in the selection of geographic areas, rather than rely on field workers to ascertain the extent of postcode boundaries, as was the case with the CCS in 2011.

**Survey Component for Overcoverage**

National Statistics Institutes (NSIs) are increasingly researching the use of administrative data for population estimates. While the majority of countries actively working towards this are intending to use a coverage survey to support the estimates, a methodology for adjusting overcoverage is yet to be established. Frameworks that are actively being worked on include methods to detect and remove erroneous records within a DSE framework (CSO Ireland)[[6]](#footnote-7), and model-based approaches to adjust for over-coverage (New Zealand)[[7]](#footnote-8). While ONS is working closely with other NSI’s to test the effectiveness of these approaches, we are also considering the potential for a separate survey component that is specifically designed to measure overcoverage for adjustment in population estimates. Following an independent methodological review of the Beyond 2011 Programme (Skinner et al, 2014), the preferred design for measuring overcoverage would be based on a dependent survey. A dependent survey involves the selection of individuals or households in the administrative records and attempting to find them in the field. These approaches broadly fall under two categories:

* *Dependent sampling:* sampling administrative records based on overcount propensity and visiting the address(es) to capture them in enumeration
* *Dependent interviewing:* sampling administrative records based on overcount propensity and

supplying field interviewers with names of individuals to verify if they still live at the address

A dependent interviewing approach is currently used by the Israeli Statistics Bureau, with a separate overcoverage sample drawn from the population register[[8]](#footnote-9). Italy have been considering a slightly different approach based on dependent interviewing, where individuals recorded on the population register that are not captured in the addresses sampled for the coverage survey receive a follow up interview to verify if they are still at the address. Currently Italy are planning to implement this survey design in October 2018.

While dependent interviewing has advantages over the dependent sampling approach, there are numerous issues that would need to be clarified in the UK context to develop an overcoverage survey along these lines. Both Italy and Israel have a system of population registers, whereas the data ONS are using in the construction of the SPD is collected by government departments for specific operational purposes. ONS is undertaking research in the following areas to support future decisions about the potential need for an over-coverage survey:

*Legal basis:* The proposed use of dependent interviewing will result in the disclosure of names of previous residents to existing residents. The legality of this needs to be understood within the provisions of the Digital Economy Act (DEA) 2017.

*Public acceptability:* ONS is in the process of conducting a series of focus groups to understand public perception towards uses of data in dependent interviewing

*Ethical Considerations:* The National Statistician’s Data Ethics Committee (NSDEC) have given advice on the ethical principles associated with this approach. An application to test the approach in October 2018 was rejected on the basis that disclosure of previous residents’ identities would breach the assurances made by ONS to safeguard personal data. Wider use of administrative data for these purposes would require agreement from the UKSA board, upon which NSDEC would reconsider the proposed use of dependent interviewing.

The dependent interviewing approaches described above will only be pursued if other methods for estimating or removing overcoverage prove unsuccessful. Failing this, there are options for developing a dependent survey that does not rely on the disclosure of previous residents to current residents, which are more in line with the dependent sampling approach referred to above. These may still involve the use of administrative records for sample selection and facilitating enumeration, but will avoid disclosing of previous residents’ identities.

**Estimation Framework**

In the absence of information that can be used as a direct measure of overcoverage on admin data, our research so far has explored the use of two estimators based on capture-recapture methods. These are dual system estimation and weighting classes estimation[[9]](#footnote-10).

In summary, the DSE approach we tested relied on the accurate detection and removal of erroneous records on the SPD. Figure 3 below demonstrates that our ‘active’ version of an SPD which removes individuals that have not recently interacted with government services, would still result in over-estimation when using a standard coverage survey adjustment with DSE.

Figure 2: DSE results using admin and coverage survey data (2011)

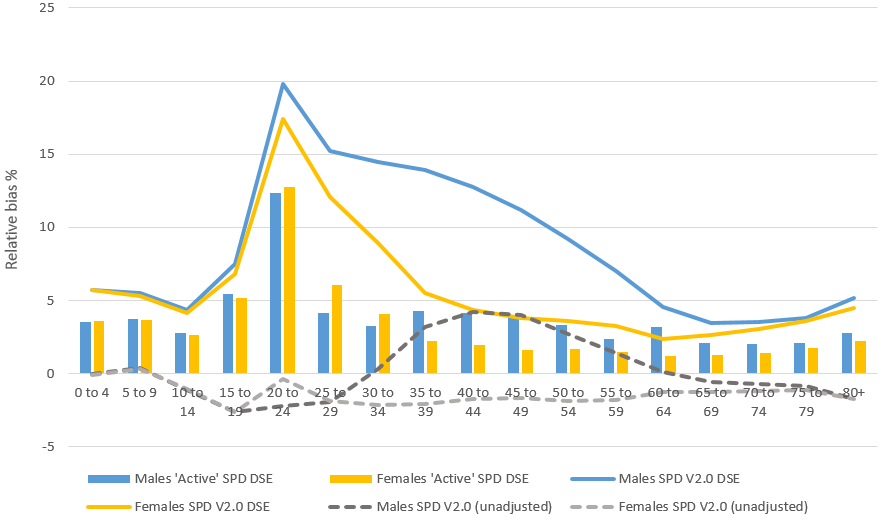
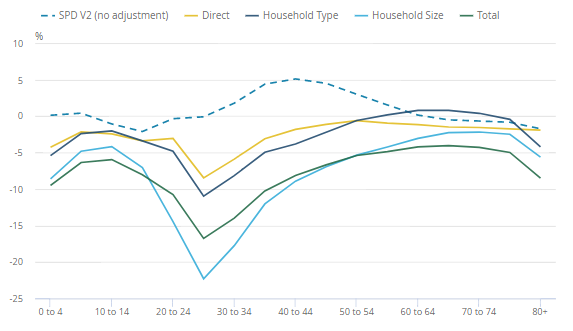


Figure 3 below shows the results of the weighting classes approach. Under this approach, addresses responding to the survey are linked to the same addresses on the admin records. Assuming that the survey has enumerated all of the residents living within the responding addresses, an adjustment weight can be derived from responding households to correct for the coverage error across all addresses on the SPD.

Figure 3: Results of Weighting Classes estimates (2011)



Different classes have been tested for the adjustment weighting, including classes based on household type and household size. In each case, the weighting classes method results in underestimation compared to 2011 Census estimates, primarily because coverage surveys are known to have within household non-response (persons in responding households that are missed in the enumeration). To account for within household non-response, an additional process needs to be put in place to measure within household non-response (for example a follow up survey), or to mitigate against it (for example dependent interviewing).

Other methods we are currently in the process of testing include model-based estimators of overcoverage (being researched in New Zealand) and iterative methods for the detection and removal of erroneous records (CSO Ireland).

**Communal Establishments**

Similar to previous censuses, a separate enumeration strategy is required to collect data on the coverage of communal establishment (CE) populations[[10]](#footnote-11) on administrative data. We have not yet developed a test plan for collecting CE data, however methods for identifying CEs for sampling purposes have started. Depending on the quality of administrative records for CE populations, we may consider the use of surveys that collect information about the number of residents and their key characteristics, rather than full enumeration of residents.

**PCS Test Plan**

A series of tests have been scheduled in the run up to 2021. Current test plans are based on the viability of integrating the PCS with the LMS:

October 2018 – A mixed mode test for online self-completion, face to face interview, telephone and paper will be collected over an 8-week period between October and December 2018. Between 12,000 and 20,000 households will be invited to take part in an integrated version of the PCS / LMS survey and the data collected will be used to test methods for estimating population size at national level for five-year age groups.

April 2019 – A similar mixed mode test is anticipated next year, with wave 2 follow up for the LMS. This may offer potential options for testing an over-coverage survey component as part of the wave 2 follow up. We will also increase sample sizes in selected areas to test methods for producing estimates of population size at local authority level.

Summer 2020 – A larger scale test of approximately 50k households sampled. By this stage we will have established whether integration of the PCS and LMS is viable and whether a separate over-coverage component is required.

Throughout the testing phase we will keep our stakeholders informed of our progress in developing the survey and estimation framework, and will publish the results of our research into estimating population size with admin data. The methods will be subject to an independent external review panel which are convening from Autumn 2018 to provide assurance on the methods being developed for the 2021 Censuses and future censuses based on administrative data.

1. *ONS Research outputs for estimating size of the population using administrative data*: [2015](https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/22october2015release), [2016](https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/researchoutputsestimatingthesizeofthepopulationinenglandandwales2016release), [2017](https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/researchoutputsestimatingthesizeofthepopulationinenglandandwales2017release) [↑](#footnote-ref-2)
2. *The 2011 Census and Coverage Adjustment Process*, ONS, July 2012 [↑](#footnote-ref-3)
3. *2011 Census Coverage Survey Summary*, ONS, July 2012 [↑](#footnote-ref-4)
4. *Predicting patterns of household non-response to the 2011 Census*, ONS, 2011 [↑](#footnote-ref-5)
5. [*ONS Labour Force Survey Background and Methodology Vol 1*](https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance) ONS, 2016 [↑](#footnote-ref-6)
6. *Trimmed dual system estimation* (Zhang & Dunne), in *Capture-Recapture Methods for the Social and Medical Sciences*, Bohning et al, 2017 [↑](#footnote-ref-7)
7. *Small domain population estimation from an administrative list subject to under and over coverage*, Graham and Lin, (unpublished) [↑](#footnote-ref-8)
8. *Evaluating the Integrated Census in Israel,* Joint UNECE/Eurostat Meeting on Population and Housing Censuses, May 2008 [↑](#footnote-ref-9)
9. [*Coverage adjusted administrative data population estimates for England and Wales, 2011*](https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/researchoutputscoverageadjustedadministrativedatapopulationestimatesforenglandandwales2011), ONS, Feb 2018

   [*Beyond 2011: Estimating the population: in theory*](http://webarchive.nationalarchives.gov.uk/20150109083021/http:/www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011--producing-population-estimates-using-administrative-data--in-theory--m8-.pdf) *,* ONS, July 2013 [↑](#footnote-ref-10)
10. *Estimation and Adjustment for Communal Establishments*, 2011 Census, ONS, Sep 2012 [↑](#footnote-ref-11)