**Estimating population size without a census**

**Executive Summary**

This paper discusses the work to date on the methodological project to estimate population size using administrative data, in the absence of a census. This project was commissioned in January 2019 by Statistical Design and Research (SDR) at ONS to develop a methodology for making population estimates using administrative (admin) data.

The aim of the project is to develop the statistical methodology to produce population size estimates and understand their quality under different assumptions of data sources, survey designs and estimators. It will do this through developing a simulation framework for admin data and survey outcomes for a base population, apply a range of estimators, and assess their performance.

The simulation begins by drawing a base population from Census 2001 data. In this first iteration of the project, the modelled area includes seven Local Authority Districts (LADs), and we include variables for age, sex, Output Area (OA), and household.

We use admin data counts from 2015/2016 to create probabilities that individuals in the base population interact with admin data. These are used to determine and flag if they are active on admin data.

Migration into and out of the modelled area is then performed, with households moved to match individual level benchmarks from 2016 migration data. We choose households at random from the base population, either to be moved out of the modelled area, or to act as donor households for those moving in.

The combination of migration and activity generate over-coverage and under-coverage in the simulated admin data, as compared to the true underlying population.

A survey of the base population is simulated to sample 1% of households, with the option of running it as a simple random sample, or as a clustered design. The simulation also allows options for altering the amount of over-coverage, and different levels of survey response, to establish twelve basic scenarios.

The resulting simulation outputs are passed to a range of estimators. This comprises some basic survey and capture-recapture estimators, for the purpose of checking that the simulation is running as expected, and a Bayesian method that could be developed as part of the estimation strategy.

By repeatedly running the survey and simulation, bias and variance measures are derived to analyse each estimators performance.

In addition to discussing the simulation model and estimators, this paper considers key assumptions, proposed improvements, and priorities for future iterations.

As part of future development, it is proposed that an underpinning programme of research which is crucial to the success of the project is undertaken. However, there are some barriers to progress, notably the challenge of migrating into a new computing environment.

**Estimating population size without a census**

**External Assurance Panel (March 2020)**

Rosalind Archer, Rob North, Aidan Metcalfe, Chris Lydiat

In this current iteration, our project involves simulating admin data outcomes, and survey outcomes, for a given “base” population. The purpose is to allow us to compare and assess estimation methods, such that we can develop a methodology for using admin data to make population estimates.

This paper represents the work that has been carried out on this project since its inception in January 2019. This project sits within the Methodology department at the Office for National Statistics (ONS), and was commissioned by Statistical Design and Research (SDR).

We have discussed this work within ONS, and we have implemented changes to reflect the feedback we received, in advance of coming to the External Assurance Panel (EAP). We would like to acknowledge and thank our colleagues for offering advice and support as we have developed this work; in particular, Patrick Graham (Stats NZ), Paul Smith (Southampton University), and Owen Abbott (ONS).

This is the first time that this project has come before the panel, and we are hoping for further feedback on the work to date, including how it can be improved and developed.

**Specific issues for discussion:**

**Q: What specific next steps do the panel think would be good for the next iteration of our project? (Section 4 describes proposed next steps)**

**Q: How do the panel think we should shape our research in the longer term (section 7)?**

**Q: Given that this project is still rather young, but that demand for insight into using admin data for estimation is quite high in ONS, do the panel have any recommendations for how this work can best be used to provide guidance to other projects in ONS?**

**Q: Are there any methods that the panel would recommend, that we have not mentioned?**

**Q: We describe several places where we believe underpinning research is required. What pieces of research would the panel recommend we undertake next, and which do they consider to be particularly important?**

**Q: How do the panel recommend that we tackle the challenges of limited access to data; are there any alternative sources of data that they would recommend we consider? – are there ways we can further capitalise on data?**

**Context and Introduction**

Using admin data for population estimation has been a top priority at ONS for some years. In April 2011, the ONS began a programme of research into whether admin data could be used to create population estimates, rather than using a Census[[1]](#footnote-1). This stream of research has developed into the current Admin Data Census (ADC) project [[2]](#footnote-2); and there now exists a significant body of work, still ongoing, which is largely carried out by the Statistical Design and Research team (SDR).

A main component of SDR’s current work is the production of Admin Based Population Estimates (ABPEs)[[3]](#footnote-3). These are datasets of individuals that are derived from linking admin data sources, and then excluding records that are believed to be for individuals who are not actually in the usual resident population.

The methodology for creating ABPEs has evolved and improved over time, but coverage problems in the admin data have proved persistent. As a result, it was considered necessary for ONS to develop a fuller methodological exploration, with the support of the Methodology department, into how to make estimates from admin data. This was the motivation for beginning the simulation/estimation project, in January 2019.

As well as instigating this stream of work, SDR have developed their research strategy into the following projects:

* Estimating “stocks” – ongoing ABPE work into population estimates
* Estimating “flows” – using admin data to estimate flows of people
* A forthcoming “hybridisation” project – the work anticipated to bring together stocks and flows into an integrated whole
* International Population and Characteristics Survey (IPACS) – a survey that will include questions to measure over and under coverage in admin data

Thus, the aims and role of the simulation/estimation project are:

* To develop a methodology for making population estimates from admin and survey data
* To ensure that this methodology can guide, and integrate with, the work ongoing in SDR – with particular reference to ABPEs, and IPACS

Given this remit we developed an initial strategy, on which our current work is based. This was to simulate admin data and a survey for a base population, in order to test and compare estimation methods. The rationale behind this strategy was:

* The simulation would create a “level playing field” on which to compare proposed estimation methods;
* A simple simulation would allow broad “stress testing”, to establish under what sort of conditions estimators do well, and under which they don’t
* If we can develop a more realistic simulation, this could be used to test and develop an admin data estimation strategy specifically for England and Wales

During our work this initial strategy has developed, and our new ideas for how to shape this project are presented later (see Section 7: Developing, longer term).

Our current approach involves an iterative process of design and implementation. The present research represents an early iteration of this process; as such, we are still relatively early on in our journey.

In this paper we present the following:

Section 1 details the current version of the simulation. We cover its design, the data that is used, and the assumptions that are readily visible. These assumptions are later addressed in Section 4.

Section 2 describes our current range of estimators for this iteration of the project. We discuss the rationale behind choosing the given selection, and we outline how we expect them to perform. We also discuss measures of performance, and what is necessary for creating a “level playing-field” for comparison.

In Section 3 we explain our choice of 12 simple “core” scenarios, which will be run for this version of the simulation and our chosen estimators. We also discuss how the random elements in the simulation can be handled in order to ensure reproducibility, and how we intend to make appropriate comparisons across results.

Section 4 addresses some of the assumptions identified in previous sections, in order to suggest improvements that could be made in the next iteration for this project. We also identify the need for a broader supporting programme of research, and suggest research questions that we believe will be valuable for the continued development of this work.

Section 5 and 6 outline two key challenges for this project, moving forward: scaling the work up in a new computing environment, and ensuring that our work integrates effectively with work elsewhere in ONS.

Finally, Section 7 reflects on how our perspective of this work has developed since this project’s initial conception, and on how we intend to take it forward.

**Section 1: Current version of the simulation – sim V2**

In this section we discuss the current version of our simulation: the design, the data, and how these two parts knit together. We also discuss how the simulation can be altered, so as to test a variety of different scenarios.

The design for simV2

The simulation/estimation project has been developed in R, outside of secure environments, for a small number of Local Authorities. These choices were made to enable speedy development, although this has also posed limitations on the data available for us to use.

The current iteration of the simulation (simV2) has the following overall design:

1. We establish a “base” population: these are people who, at time T0, are in our population of interest (“LAD7” – see below)
2. At T0 we allow people to have an “activity event” – i.e. they have a chance to appear on a range of admin data sources
3. We allow people to move in and move out of the population, between T0 and T1
4. At T1 we carry out a survey
5. The resulting survey and admin data outputs are passed to a selection of estimators, for making population estimates (see Section 2. Current range of estimators)

In this design, over-coverage and under-coverage are thus determined by people moving and interacting with admin services (Figure 1). We note that we do not currently move people within the modelled area – this is a refinement that we intend to implement in our next version (see Section 4. Development: Making improvements, and the next version).



Figure 1: A diagram presenting migration and activity events across time

To inform the design of simV2, we use the following data:

* The base population is drawn from non-identifiable Census 2001 data, for an Estimation Area that comprises 7 LADs in Lincolnshire (Boston, East Lindsey, Lincoln, North Kesteven, South Holland, South Kesteven, West Lindsey)
* Published migration data is used for moving people into and out of the population. Migration is broken into two steps to accommodate our data:
	+ For migrating people between LAD7 and anywhere that is outside of E&W, we use long-term international migration data, for the year up to mid 2016
		- This data is stratified by sex and broad age groups (0-15, 16-24, 25-34, 35-44, 45-54, 55-64, 65 and over)
	+ For moving people between LAD7 and anywhere within E&W, we use national (internal) migration data, also for the year up to mid 2016
		- This data is stratified by sex and ten-year age groups (0 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70, 71 and over)
* Admin data counts for 2016 are used for deciding whether people in the base population have an activity event at T0 (i.e. interact with admin services)
	+ This data was made available for 6 admin sources: births, benefits, PAYE, HESA, PR[[4]](#footnote-4), and School Census
	+ Admin data counts are aggregated according to age, sex, and LAD;
	+ This data is for “active” counts, meaning that before we received it from the team in Statistical Design and Research (SDR) it had exclusion rules applied to take out “inactive” records

For further details about the data, and how it is prepared for use in the simulation, please see Appendix (1. Data preparation details).

Bringing together data and design:

We now describe in more detail the three main mechanisms that are used in the current simulation:

* How people are allowed to have an activity event
* How people are moved
* How the survey is simulated

How people are allowed to have an activity event

For each admin source, admin counts per age/sex/LAD stratum are compared against Mid-Year Estimates (MYEs) per stratum. We use these ratios to make probabilities for whether someone in the base population with a particular age/sex/LAD combination will interact with each admin source. For those strata where the number of admin counts is larger than the MYE, we cap the resulting probability at 1.

We then use these probabilities to simulate whether an individual in the base interacts with each admin data source. The output is a 0/1 flag for each individual that indicates whether they are on a source at T0.

This design implies several important assumptions:

1. An individual’s admin data presence across admin sources (“admin data footprint”) is established at T0, and is assumed to stay the same for T1 – there are no further admin events
2. Immigrants are not allowed to have activity events, as they arrive after T0
3. We have perfect linkage, and no errors, missingness, or disagreement between sources in the admin data: each individual has a series of 0/1 flags for activity across sources, and this information is assumed to be correct and full
4. An individual’s presence on a given admin source is independent of their presence on any other admin source (independence between sources)
5. An individual’s presence on a given admin source is independent of whether or not any other individual is on that source (independence between individuals)

We shall discuss some of these assumptions later, in terms of what we hope to improve in our next simulation version (see Developing the simulation).

How people are moved

In simV2, people are moved by household, where possible. We chose to do this in order to maintain the household ID variable in the data at T1, and so allow household level sampling in the simulated survey (see How the survey is simulated, below).

The migration data provides benchmarks for how many individuals in total we wish to move, across age-sex strata. Households are then chosen at random from the base population to meet these benchmarks. So, for moving people out, we choose households from LAD7 to flag as “out” flow; and for moving people in, we choose households from LAD7 to copy and add to the population at T1 as “in” flow. Those immigrating into LAD7 are given fresh household IDs, to disambiguate them from the donor households.

In practice, the process of moving households to match the “individual benchmarks” is straightforward when the stratification for moving people is quite broad – as for international migration. However, this process can become problematic when the stratification for the benchmarks becomes more granular, or when there is too much variation between the number of people in each stratum. Effectively, the code is designed to accept households so long as individual benchmarks are not over-filled, but some age/sex/LAD strata fill up faster than others.

We developed two generic working solutions to help the code to run:

* Lower the requirement that each stratum benchmark needs to be fully met (e.g., require a 0.8 threshold)
* Allow households to be sampled with replacement

This “household allocation” code could certainly be improved, and for our next version of the simulation we would like to explore Combinatorial Optimisation methods for a potential solution. Developing a flexible and consistent method for moving people is important, as we expect that this will become more difficult if we wish to run the code on areas that are different to LAD7 (e.g. London), or include migration within the chosen area (i.e. intra-LA moves).

Similarly, we would like to develop the code to ensure that robustness to specific pitfalls is “built-in”. For example, in the current code, when we sampled donor households with replacement, we did not observe any particular household being used as a donor a large number of times; however, we wish to build a limit for this into future code, in order to safeguard against this possibility.

For migration LAD7 and the rest of England and Wales, we had to include a further specific alteration to accommodate the large number of young people moving into and out of Lincolnshire (which we believe to be due to the presence of Lincoln University). This involved moving a tranche of young people as individuals, before applying our algorithm for household allocation.

To do this, we capped the number of moves across age/LAD strata at the 75th quantile, for young people (18 – 25 years old). This split our migration benchmarks into a “high moves” dataset, and a dataset for household allocations. We moved households first, and then chose individuals to satisfy the remaining “high moves”. This meant that some households in T0 had a new structure in T1, where a student-aged person had been moved away, but the rest of the household had not. It also meant that a tranche of young people were moved into LAD7 as individuals – for expediency, they were moved in as single-person households.

We note that this method of moving people creates the following further assumptions in the simulation:

1. Using the base population as donors for migrating households assumes that the distribution of household structures in the base population reflects that of the population immigrating into LAD7, as well as that of the population that is leaving
2. Households in the base population do not alter in their composition between T0 and T1 – except where we have had to move individuals to account for the young “high peak”. We think it would be more realistic to allow for further instability in household structure in the base population (i.e. households change over time).
3. A more specific application of point g.: Although we allowed students (student-aged people) to leave LAD7, and allowed some to arrive from outside, we didn’t account for those in the base population who would leave home to attend university locally – i.e. removing young people from base population households, and allowing them to re-appear elsewhere in LAD7
4. It is highly unlikely that all student-aged people arriving in LAD7 would appear in the data as single-person households; it would be more realistic to consider that they join communal establishments.

How the survey is selected:

The survey is selected from the population at T1: that is, from those who were in the base population at T0 and did not emigrate out of LAD7, and those who immigrated into LAD7.

The survey units are households, and we chose a survey size that is of 1% of the population, to reflect current thinking for IPACS. We allowed for non-response to be simulated at both household and individual level (see below), so that after the survey step the data gains two 0/1 flags: one for every household’s response, and one for every individual’s response.

In addition to those above, the design of the survey makes some further assumptions:

1. We are effectively conflating household and address by using household ID as our sampling unit
2. The survey is conducted at a single point in time – this is not realistic, given that IPACS will be run in a series of waves

Creating options in the design

In simV2, we wished to establish some simple, changeable, settings in the simulation, such that we could create a variety of “scenarios”. The purpose of these scenarios is threefold:

* Provide a range of conditions that can demonstrate that the simulation is operating as one would expect
* Test the behaviour of estimators
* Create a core set of scenarios, intended as a reasonable starting point for further hypothesis-creation

To enable this, we created the following options for running the simulation:

1. Altering the survey structure
* We coded two very simple options for how to choose a survey:
	+ A simple random sample, where we choose 1% of households at T1; or
	+ A stratified clustered design, where in each LA 1% of OAs are chosen, and every household is sampled within those OAs.
1. Altering survey response
* We allowed three types of response pattern:
	+ Full household and individual response
	+ Household non-response, but full individual response within responding households
	+ Both household non-response, and individual non-response within responding households

We are aware that the situation in which household non-response might arise, but not individual non-response, is somewhat implausible. Our rationale in doing this was to build up non-response in layers, so that we can think about how the two types of non-response differ in their effects, and so lay the ground-work for a better simulation.

1. Altering over-coverage in the admin data

In trying to create an option where we can increase over-coverage we encountered two questions: how to increase it, and by how much?

Altering over-coverage in sim V2 was a little trickier than anticipated, because it is determined by migration and activity alone. It was also unclear as to how much we ought to increase over-coverage in order to create a “high over-coverage” option. Essentially, we want to increase it enough for over-coverage to become noticeably problematic for estimators, but we also recognise that we could use this opportunity to benchmark the over-coverage to known admin data.

Our first attempt to increase over-coverage was by doubling emigration; however, this led to moving so many people out that the population at T1 was smaller than that at T0, which seemed implausible. We then considered increasing the probabilities that individuals interact with admin data at T0; in contrast to our first attempt, this approach meant that the population size at T1 is not affected, but we found that the degree to which over-coverage can be increased is limited.

Currently, we are testing the option of creating duplicates in the admin data as a way to further increase over-coverage. In the future, we hope to have greater control over how coverage problems in the simulated admin data by improving the design – for example, by including data linkage error, or extending the simulation over a longer time period, so that coverage errors can accrue.

As a result of these explorations, and discussion in ONS, we now understand that the simulated admin data is much too clean and accurate, and that a key piece of work will be to better reflect the degree of over-coverage and under-coverage that is known to afflict real-world admin data (see Section 4. Development: Making improvements, and the next version – Making messy admin data).

These option in the design are all rather crude, but they gave us a basis on which to introduce differences in survey type, response pattern, and admin data coverage. We chose to alter these particular design elements to help us test the simulation, as part of a “core” set of 12 scenarios (see Section 3: Running the simulation and estimators).

Once the simulation has been run, and we have obtained outputs for admin data presence and survey response, we can move to the next part of the process: running estimators.

**Section 2. Current range of estimators**

At present, our chosen set of estimators include: Horvitz-Thompson, Ratio, DSE, a two-stage estimator, and two versions of a model-based Bayesian estimator devised by Patrick Graham at Stats NZ.[[5]](#footnote-5)

These estimators can be considered in two groups, according to their purpose in our project:

1. Those that are included for comparison purposes, and for sense-checking the simulation (Horvitz-Thompson, Ratio, DSE, two-stage: “comparator estimators”)
2. Estimators that we believe could be developed as part of an admin data estimation methodology (Bayesian PG estimators: “candidate estimators”)

We have also identified other possible candidate estimators that we wish to explore in the future (see Section 4. Development: Making improvements, and the next version).

Comparator estimators

These estimators were chosen for comparison purposes with candidate estimators, and for sense-checking the simulation. We did not make any special alterations to these methods, using very simple versions of each. The two-stage estimator involved applying the Horvitz-Thompson estimator for households, per OA, then using the Ratio estimator to combine these OA-level estimates up to the level of LAD7. Further details regarding how these estimators were calculated can be found in the Appendix (2. Comparator estimators).

Our comparator estimators influenced our choice of scenarios in the following way:

* We expect HT estimation to struggle in the face of non-response, unless we are able to include an appropriate adjustment factor
* We expect Ratio estimation to fare better than HT estimation when it comes to non-response, providing that the correlation between admin data and survey responses is good, per estimation stratum
* We expect DSE to deal reasonably with non-response, but to struggle with over-coverage
* We expect the two-stage estimator to show improved performance with the more complicated survey design and perform poorly with a simple random sampling design

We are aware that our current choice of scenarios and estimators could be expanded and explored; the current choice is simply intended as a first step.

Adjusting for non-response

We chose to allow our estimators to benefit from an adjustment weight for household non-response, but not for individual non-response. The rationale behind this was that it seems plausible that one might be able to gather information regarding the total number of units in a sample that do not respond (household non-response), but that it is much less plausible that one would be able to measure individual non-response.

The Horvitz-Thompson estimator, and the Patrick Graham estimators, are adjusted to deal with household non-response, where this arises.

Currently, for convenience, we have modelled response rates as flat across individuals. In future iterations of the simulation we hope to update the probabilities we use for individual non-response, so that they are more realistic.

Candidate estimators

For this version of the simulation we included two versions of Patrick Graham’s estimator, as described in examples in his 2018 paper[[6]](#footnote-6). These two methods have been dubbed the “full Gibbs” approach, and the “back-calculation” approach.

Essentially, the first is a fully Bayesian treatment of the problem as described and parameterised in Patrick Graham’s work. In comparison, the second makes a shortcut, by not modelling every parameter explicitly; instead, a key parameter is calculated in a deterministic fashion from other parameters that are modelled directly (this is the “back-calculation”).

The Gibbs version is more computationally expensive, but we believe it is more reliable than the back-calculation, especially when the estimation task becomes more complex – that is, it moves beyond estimates that are stratified by more than a single binary variable, and beyond sampling designs that are more complicated than simple random sampling. Furthermore, as the Gibbs sampler is fully Bayesian it not only provides meaningful credible intervals, but also the flexibility to incorporate more complicated aspects – for example, different survey designs and hierarchical modelling to allow pooling across covariates.

A drawback of the Full Gibbs approach is that it is not easily parallelised, whereas this is trivial for the back-calculation. This is a concern for us, because we expect to move into a new distributed computing environment in the future (see Section 5. A key challenge: Scaling up).

A third version of Patrick Graham’s estimator is a “conditional likelihood” approach, which may provide a happy medium between Full Gibbs and back-calculation. This method is less computationally burdensome than a Full Gibbs approach, and as it relies on Monte Carlo (MC) algorithms, rather than Markov Chain Monte Carlo (MCMC), we expect it to be vastly easier to parallelise[[7]](#footnote-7). Also, it does not suffer from some of the known drawbacks of the back-calculation approach. Currently, this is in development, and we anticipate adding it to our list of candidate estimators.

A known challenge for this estimation method, across all versions, is that it is very sensitive in terms of the “lambda” parameter – this is the capture probability for an individual, given their characteristics. To run this method, lambda must be supplied at the outset; when it is supplied incorrectly, we expect the results to be affected. For this reason we are planning to conduct sensitivity testing, to see to what extent the results from this method are affected by deviations in the value of lambda supplied, and the true value.

For these models, we have used non-informative priors, as recommended by Patrick Graham.

In terms of our core scenarios, we expected our candidate estimators to:

* Deal well with over-coverage
* Struggle when the survey was more complicated than a simple random sample (as this will cause the true value of lambda to differ from the supposed value)
* Struggle when there is non-response (as this will also cause the true value of lambda to differ from the supposed value, except if adjustment can be made)

We note that our current implementation for this method is still extremely basic. There is work that could be done to explore ways in which to refine and fit this model (see Developing estimators).

Measures of Performance

In the current version, we have chosen some basic “measures of performance” by which to compare estimators: bias, relative bias, variance, Mean Squared Error, Relative Mean Squared Error, and Relative Root Mean Squared Error.

These measures are based on point estimates, and on their variance over the runs that are carried out per scenario. For the Bayesian models, we use the mean posterior value as a point estimate, and variance is calculated from how mean posterior values vary over runs, per scenario.

We recognise that although we want to compare estimators, this is not straightforward. Using the above measures of performance is useful, but Bayesian estimators provide more information than the other estimators, in the form of complete distributions.

So for these models we intend to generate further measures per run: credible intervals, and the % coverage of credible intervals. We also calculate and examine the Gelman Rubin statistic, in order to assess convergence. We chose these measures because they were used by Patrick Graham to analyse his results, and they are recommended by Lambert et al in their paper exploring simulation and Bayesian models.[[8]](#footnote-8)

Can we ensure a level playing-field?

More generally, we are aware that we will face an increasing challenge in testing and comparing estimators, as we expand our range of estimators. How do we ensure that we do, indeed, have a “level playing-field”?

This is partly assured by passing the same dataset to each estimator during a given run/scenario; however, there are at least three further considerations:

1. Model-based estimators need to be fitted to a given problem; a poorly-fitting model will not perform as well as a well-fitted one. We do not wish to unfairly reject a candidate estimator when it has not been fitted as well as a competing estimator.
2. The outputs from different types of estimator can be difficult to compare, and there may not be an obvious way to do this; for example, Bayesian model outputs are quite different to those from design-based estimators. In such situations, we expect that the performance criteria for estimators may not overlap, making direct comparison difficult or impossible.
3. Similarly, when there are multiple criteria on which to judge estimators, we expect that we will need to take into account the performance of estimators as a whole.

It is likely that when we come to compare more estimators, the comparison process will become a more qualitative assessment that takes into a consideration a wide range of measures of performance, and balances these against a wide range of criteria (e.g. ease of interpretation).

**Section 3: Running the simulation and estimators**

Simulation V2 is the first iteration of our simulation to be fully versioned: that is, whilst some scenarios were run for V1, this was done in an ad hoc manner. For V2, we have spent further time and effort in co-ordinating the code, in checking its quality, and in choosing and running a set of basic scenarios.

We chose 12 simple “core” scenarios to test the simulation. Our aim was:

* to show that the simulation could run end-to-end from starting data to suitable measures of performance, and see if it was functioning as expected
* to identify and test hypotheses for what would stretch our selection of estimators,
* to identify assumptions and create further hypotheses

Essentially, by following the simulation through to a set of results we have learned a lot about how to knit our code together into a consistent whole, and we have begun to think about what the simulation can usefully do.

Scenarios

The twelve core scenarios test different combinations of: over-coverage in the admin data, sampling strategy in the survey, and types of non-response. The full list of these scenarios are found in the table below (Table 1).

|  |  |  |  |
| --- | --- | --- | --- |
| **Scenario** | **Amount of overcoverage** | **Coverage survey response** | **Sampling design** |
| 1 | Default | All households respond | Simple random sample |
| 2 | Default | 50% household response | Simple random sample |
| 3 | Default | 50% household response, individual non-response | Simple random sample |
| 4 | Increase | All households respond | Simple random sample |
| 5 | Increase | 50% household response | Simple random sample  |
| 6 | Increase | 50% household response, individual non-response | Simple random sample |
| 7 | Default | All households respond | Stratified clustered design  |
| 8 | Default | 50% household response | Stratified clustered design  |
| 9 | Default | 50% household response, individual non-response | Stratified clustered design  |
| 10 | Increase | All households respond | Stratified clustered design  |
| 11 | Increase | 50% household response | Stratified clustered design  |
| 12 | Increase | 50% household response, individual non-response | Stratified clustered design  |

Table 1: Chosen scenarios for running simV2

For each scenario we decided to carry out 100 runs. Ideally, we would like to run more, but we chose relatively few in order to create a manageable workload for initial testing.

Reproducibility

We coded a function to provide seeds throughout the simulation. This function draws seeds from an initial pool, without replacement, and keeps a record of which seeds are used per run. This produces a fixed list of seeds to allow both reproducibility and comparison across scenarios.

Throughout the code, there are various places where random draws are required. Seeds are required at the following points in the code:

1. To select which households are moved out of the base population, or are used as donor households for the immigrating population. Every time a household is selected by the household allocation algorithm – whether or not it is accepted in the allocation – a seed is used.
2. To select which households are sampled in the survey
3. To select which households respond to the survey
4. To select which individuals respond to the survey (amongst responding households)

In total, about 30,000 seeds are required to run through the current simulation. These are drawn from a pool of 100,000 seeds.

Comparisons

Not every result from the simulation is comparable with every other result. Results across estimators may be compared for a given scenario, as the same population per run is passed through to all of the estimators.

Similarly, we are able to set seed such that in every scenario, the ith run will begin in the same place. However, looking at the ith run across all scenarios, the resulting populations will differ, depending on at what point any two scenarios part company. This is represented in the data flow below (Figure 2).



Figure 2: A diagram respresenting the flow of data from initial base population to final simulated population, to make 12 core scenarios

So, for example, on the ith run:

* scenario 1 will differ from scenario 4 in terms of those included in the survey, but they will have the same individuals in their over-coverage and under-coverage, as the admin data activity and migration steps occur upstream of the survey
* in comparison, scenario 1 will differ from scenario 7 both in terms of the survey, and in terms of over-coverage and under-coverage

We also note that scenarios 1, 2, and 3 are effectively nested. That is, in the code, household non-response is applied to the full survey generated for scenario 1, to arrive at scenario 2; and individual non-response is then applied to arrive at scenario 3. Similarly, scenarios 4-6, 7-9, and 10-12 are nested.

The intention behind this design is for efficiency, cutting down on computation time; it also allows us to weed out the impact of one scenario over another by reducing the variability between them.

Running the simulation and estimatorsow

It takes approximately an hour to produce a single run through all scenarios. The lengthiest parts of the process are the migration steps, and running the Bayesian Gibbs estimator. As we seek to include more covariates, we expect the runtime to increase exponentially.

Although we have identified some coding improvements that could decrease the time required to run the simulation, the question of how to expand the simulation and estimators to a larger scale remains outstanding. We believe that one route to making this possible is to take advantage of distributed computing in the new Data Access Platform at ONS (see Section 5. A key challenge: Scaling up).

For our current version we have developed our code to produce population estimates for the total population, and the total population stratified by gender. We are currently working on expanding the estimators to include age group as a covariate; this seems particularly important, as we are aware that admin data activity patterns are highly variable over age.

Establishing good practice

We are mindful that we are yet to establish a full set of processes to assess how the simulation is performing. In particular, we are aware that randomness is brought into the simulation at several points, and we want to understand how this uncertainty manifests in the simulation outputs.

Our current thoughts around how we might investigate this are:

* to investigate how random different outputs from the simulation are, and whether they are prone to systematic biases (e.g. distribution of true population sizes, distribution of over-coverage, distribution of characteristics in different outputs, etc)
* to identify and investigate extreme and outlying values in both the simulation outputs and in estimates
* to identify good comparison points in the simulation, where we can compare or benchmark an output with data (e.g. known age-sex distributions in admin data)
* descriptive statistics for outputs per run, so that we can identify runs that have extreme or outlying values in the population

Results

Following an internal review at ONS with the Census Research Assurance Group (CRAG), we decided to incorporate some of the advice we received, and to fix a couple of known bugs. This has delayed the production of our most recent results. We shall be providing a set of initial results during next week (24th- 28th February), and we thank the panel for their patience in this matter.

**Section 4. Development: Making improvements, and the next version**

We have already identified several limitations and assumptions in the existing project. In this section we will outline potential improvements that could be carried out, in terms of design and data. We also discuss supporting research into admin data, which we believe is necessary to enable development of the simulation; and we discuss new candidate estimators that we would like to explore. Lastly, we propose the next best steps for us to take in the next iteration of this work.

Design improvements

Building sim V2 of highlighted many assumptions that we would like to improve upon. Taking these one-by-one, we suggest some possible improvements:

1. An individual’s admin data presence across admin sources (“admin data footprint”) is established at T0, and is assumed to stay the same for T1 – there are no further admin events
2. Immigrants are not allowed to have activity events, as they arrive after T0; We also do not allow people to have more than one migration event

Suggested improvement: We think that we can address this as part of extending the simulation over time. Our idea is that we would change from modelling time as continuous, rather than being represented by two discrete time points. In this framework, we would model “time-to-event” for activity and mobility. This would be need to be supported by relevant data regarding the frequency with which different types of individuals interact with each admin source.

1. We have perfect linkage, and no errors, missingness, or disagreement between sources in the admin data: each individual has a series of 0/1 flags for activity across sources, and this information is assumed to be correct and full

Suggested improvement: We are considering including a linkage error step in the simulation by perturbing the current perfectly-linked dataset. This could include inducing false positive matches, and false negatives; ideally these would be based on information regarding false positive and false negative rates.

1. An individual’s presence on a given admin source is independent of their presence on any other admin source (independence between sources)
2. An individual’s presence on a given admin source is independent of whether or not any other individual is on that source (independence between individuals)

Suggested improvement: These assumptions suggest that we require further research into the dependency patterns in admin data. After some discussion at our internal Census Research Assurance Group (CRAG) in January, we understand that Multiple System Estimation could be a good candidate method for unpicking these patterns (see Section 5. A key challenge: Scaling up).

It seems plausible to us that such patterns exist, and are related to a person’s life-course. For example, as people get older, they would disappear from the School Census, and may appear on HESA or PAYE data. Similarly, it seems likely that mobility and activity are also related to a person’s life-course.

1. Using the base population as donors for migrating households assumes that the distribution of household structures in the base population reflects that of the population immigrating into LAD7, as well as that of the population that is leaving

Suggested improvement: it seems likely that immigrating households are different in composition to the households that are currently in an area – particularly in terms of international immigration. Given that there is a broadly consistent pattern in household type distribution for migrating households, we could use data from the Census to inform this intuition (variable for “address one year ago”).

In addition to this, it is worth considering that particular areas may attract certain types of household – for example, catchment areas for schools, areas for student private rentals, and areas with particular minority groups. Potentially we could expand the simulation to take advantage of this kind of pattern, too.

1. Households in the base population do not alter in their composition between T0 and T1 – except where we have had to move individuals to account for the young “high peak”. We think it would be more realistic to allow for further instability in household structure in the base population (i.e. households change over time).
2. A more specific application of point g: Although we allowed students (student-aged people) to leave LAD7, and allowed some to arrive from outside, we didn’t account for those in the base population who would leave home to attend university locally – i.e. removing young people from base population households, and allowing them to re-appear elsewhere in LAD7
3. It is highly unlikely that all student-aged people arriving in LAD7 would appear in the data as single-person households; it would be more realistic to consider that they join communal establishments.

Proposed improvement: These assumptions all speak to the challenge of reflecting household (in)stability in the simulation. We expect that the design could be improved to reflect plausible reasons for household change – for example, marriages, civil partnerships, divorces, births, deaths, and students leaving home. Similarly, there are datasets that could, in theory, be used to inform the simulation – such as the Labour Force Survey, Annual Population Survey, Deaths register, and births register.

However, we are aware that household stability represents a difficult modelling problem, and that arriving at a solution is not facile; for now, we note that other teams in ONS have experience in estimating households, and that other projects are exploring how changes in household structure manifest in admin data. In developing our own approach, we would like to sync up with, and benefit from, these other avenues of research.

We also note that household stability relates to the question of dependency between individuals appearing in admin data (assumption e.).

Lastly, the current simulation already makes some crude provision for moving students (as a “high peak”). This deviation from our household allocation method arose from necessity, to allow the design and data to fit together, and so seems to be a nice indication that the strategy of improving both design and data usage in the simulation naturally brings to light important assumptions and mechanisms.

1. We are effectively conflating household and address by using household ID as our sampling unit
2. The survey is conducted at a single point in time – this is not realistic, given that IPACS will be run in a series of waves

We currently do not have a proposed improvement for assumption j, however, this is something that may become more tractable as we gain familiarity with IPACS (see Section 6. A key challenge: Integrating our work with SDR). Assumption k could be addressed if we were to develop our simulation to being continuous over time.

Further design improvements

In addition to the improvements mentioned so far, in the next version we would like to expand the design in the following ways:

* Include movements between LADs in LAD7 (internal movements). This would allow us to recreate the characteristic coverage problem whereby someone moving but not interacting with admin can create both under-coverage and over-coverage at the same time.
* Testing different geographies – for example, running our existing code on Birmingham (a large, urban area).

Data improvements

A key part of developing the simulation is to improve not just the design, but also our data usage. This includes finding new data to support the design, and making sure that the data we are using is included appropriately. Specific places where we would like to improve the data we are using include:

Base population: we have two plans to improve this data:

* In the future, use Census 2011 data, linked to admin data. This will require us to work inside DAP, which is where this data will be securely stored (see Section 5. A key challenge: Scaling up)
* In the short term, inflate our BASE population at T0, so that it is approximately the same size, and has the same demographic distribution across age and sex, as the population for LAD7 in 2015/2016. This will address some of our concerns about using 2001 Census for the base, while using admin counts and migration data for 2015/2016.

Migration data: we have used published data for migration. We are still considering if we are using the best available data, and whether we are implementing it in the right way; this is a difficult question, as we have discovered migration data to be very complex. We note that currently we are using the admin data that lies behind ABPEs (i.e. for “stocks”), and that admin data also exists for “flows”; we would like to include this in our simulation as well (see Section 6. A key challenge: Integrating our work with SDR).

Admin data: As mentioned previously, admin data are gathered in a continuous way, and the various sources do not necessarily all refer to the same collection window in time. We would like to expand our work to potentially reflect this aspect of admin data. We also wish to take advantage of admin data concerning records for “inactive” people, as described in Section 1: Current version of the simulation – i.e. those who did not have an activity event during the given window, but are present on the admin source nonetheless.

A further consideration around how we use data in the simulation, is that we wish to avoid datasets being over-used. For example, the published data being used currently for national migration is derived from admin data.[[9]](#footnote-9) Given that various published statistics can be based on the same collection of data sources, it is possible that we could inadvertently re-use the same sources in an inappropriate way (e.g. validation). Similarly, we will need to be careful about the definitions being used in the simulation and in the data being utilised; it is commonplace for admin data definitions to differ across sources, or to be different in comparison to Census and survey variables.

Research to inform design and data

Ideally, we would like to develop our design and data usage together. We have found that useful information about our admin data can be difficult to find, so we expect that in order to continue developing this work we will need to also develop a programme of supporting research. Currently, we expect that this will be largely undertaken by Statistical Design and Research (SDR), for reasons of resource and of knowledge: this team has by far the greater experience of dealing with admin data.

We have proposed a set of research questions to SDR as being useful ways to increase our understanding of admin data such that we can improve this project. These can be found in the Appendix (3. Proposed research questions).

Making messy admin data

A general comment we have received about our current work is that the “combined admin data” we simulate is too good, and lacks the range and intensity of coverage problems that we know to be in our real-world admin data. Thus, one of our priorities going forward is to simulate messier admin data.

To do this, we believe that the following changes would help:

* Design improvements:
	+ Extending the simulation over time, from discrete to continuous (as mentioned above), would allow errors in the simulated admin data to compound over time
	+ Introducing linkage error, between admin sources, and between admin and survey
* Data improvements:
	+ Make use of the “inactive” admin counts data, which were excluded from our current admin counts data (see above)
	+ Benchmark over-coverage and under-coverage in our simulated combined admin data to reflect the observed coverage problems in previously-produced Statistical Population Databases (SPDs) and Admin Based Population Estimates (ABPEs) – the population estimates produced by SDR

An important consideration in improving simulated admin data is that we wish to recreate the messiness that is characteristic of admin data for England and Wales, rather than messiness per se. In order to do this, we would need to carry out supporting research into the admin data available to us – for example, into linkage error rates between given sources.

More generally, we would like to expand the way that admin data is produced and used in the simulation. That is, instead of simulating a single “combined admin” dataset, we want to develop a range of approaches for using admin data. This could include using single sources, building up to a “family” of combined admin datasets, whose members would differ according to the strategy used to create them. For example, one strategy might be to remove any record that is not of the highest standard (e.g. with respect to activity, linkage, error), with the aim of removing over-coverage entirely.

As part of expanding how we simulate admin data, we are thinking about benchmarking over-coverage and under-coverage to known SPDs and ABPEs. Combined admin datasets based on this approach would thus reflect the processes that SDR have used to combine admin data sources and remove non-resident records.

This development would not only provide an indication of “how much” over-coverage should be in a combined admin data source, it may also help us to align our work more closely to that which is ongoing in SDR. (for further discussion, see Section 6. A key challenge: Integrating our work with SDR).

Developing estimators

As already mentioned, the Bayesian Patrick Graham estimators are known to be problematic in terms of the “lambda” probabilities for survey capture. We are currently investigating this, exploring the impact of this on the results that are generated, and whether we can devise a solution.

However, we believe that this methodology has the potential to be developed, and for this reason, we would like to continue working with it. In particular, we expect that it can be usefully altered to more flexibly model the probabilities for whether individuals are sited in the true population and/or captured in the admin (the phi parameters). We are also interested to see how this approach performs against our comparator estimators in more complicated scenarios.

We also note that this method has subsequently been developed more fully in Stats NZ, to account for data linkage error (given data from a clerical matching exercise), and for misclassification in admin records.

Besides continuing our investigation into Patrick Graham’s methodology, we hope to explore two other estimation approaches. These are:

* A second Bayesian methodology that has been developed by John Bryant and Junni Zhang (Bayesian Demographic Accounts)
* Latent class analysis, where the latent variable codes for whether someone is in the true underlying population

When applied appropriately, Bayesian Demographic Accounting appears to accommodate a variety of data sources in a single model, without the need for linkage. It also allows one to include demographically plausible assumptions in the model, such as steady population growth; and it can take advantage of data sets available across time.

Similarly, we believe that a latent class analysis may be adaptable to data over time. Such approaches could be helpful for estimating from data that is produced over time, such as admin data and IPACS.

In setting out our plans for development, it is worth mentioning that the degree to which we can develop and test estimators depends upon our simulation being sufficiently complex. That is, we require the simulation to offer a sufficiently sophisticated and realistic problem, such that we can test and assess the benefits of the different models. For example, at present, our simulation only describes coverage patterns in terms of age, sex, and LAD; thus, it is not particularly useful to include further covariates in the estimators – we cannot supply them with a simulated reality in which extra covariates will give them an edge.

In the light of the above, potential improvements for the next iteration of our project are:

* Include internal movements, between LADs in LAD7
* Extend the simulation over time, and model activity and mobility as time-to-event
* Develop how admin data is simulated: single sources and a “family” of combined admin datasets (see also Section 6. A key challenge: Integrating our work with SDR)
* Include mechanisms for data linkage error, in a basic fashion
* Explore Multiple System Estimation, to unpick the patterns of dependency between-source and between-individuals in the admin data (this depends rather heavily on getting access to appropriate admin data – see Section 5. A key challenge: Scaling up)
* Improve use of data: e.g. more comparison to data to ensure that outputs across the simulation are realistic
* Test our code with different geographies
* Expand our analysis of simulation outputs

In the following section we will extend our discussion about how the project should develop to describe two key challenges that may require us to alter our direction more significantly, and so require special consideration: scaling up, and integrating this project with SDR’s existing work.

**Section 5. A key challenge: Scaling up**

We intend to continue and develop our project in DAP, the new secure environment in ONS. However, moving to this new environment is not trivial, and involves several key challenges. In this section we describe these challenges, as well as our thoughts about how to scale up our work.

What is DAP?

DAP is a new secure environment in ONS, built using Scala, and enabling distributed computing. It is intended that DAP will replace some of our core secure environments, and we anticipate that most of the supporting data for this project will be stored here - both Census 2011 (and 2021) data, and admin data.

We also anticipate that that is where the “demographic index” will be made available to researchers. This is a linked dataset of Census and admin data records that is currently being constructed in ONS. It involves matching records “in the clear”, such that we expect the linkage to be of better quality than has hitherto been possible. Once it has been successfully linked, the data will be “de-identified”, and made available to analysts at ONS. Furthermore, the scale of this linkage exercise is much larger than anything attempted up to now, including linkage over time as well as over source, so we hope that this will open up new and better avenues for research.

Why is moving to DAP difficult?

Moving to DAP requires over-coming several difficult steps. Some of these are common to all projects in ONS, and some are specific to this project.

Firstly, DAP is a work in progress. Although the platform is up and running, the data is still being moved in, and the processes for enabling access to researchers are not fully established. Also, since many projects are being moved over to DAP, there is a substantial administrative overhead: this is a lengthy business and will have some impact on how quickly we can expect our project to develop.

Even when data are available, researchers may not necessarily have access – as part of ONS’ commitment to data providers, there are restrictions on who may have access. The trickiest sticking point for this project is a general rule that linkage must be uncoupled from analysis; this is another main reason for why we wish to use the demographic index.

Secondly, although researchers who access DAP can make use of sessions that have up to 16GB RAM, it is anticipated that large-scale work will involve processing in a distributed fashion. DAP is an idiosyncratic coding environment, and in order to use distributed computing we will need to use Spark. This task can be made more tractable by using Python or R as an intermediary language (via PySpark or sparklyR), however, there remains a substantial translation and re-interpretation task of existing code – effectively, a steep learning curve.

Thirdly, quite apart from learning how to deal with new coding languages, distributed computing poses significant problems for some types of analysis. Essentially, distributed computing can cut down the computational strain of a process by splitting the workload amongst multiple “worker” computers; however, this requires that the process can be split into parts, or “parallelized”.

Unfortunately, some types of analysis are not easily parallelizable – for example, Bayesian analysis. Individual chains can be passed to workers, but each chain needs to be run in series. We have been investigating some potential solutions to this problem, but there is no readily available solution, and it remains a problem to the wider research community beyond ONS.

**Section 6. A key challenge: Integrating our work with SDR**

An important role for our project is that it should be able to work alongside the work in Statistical Research and Design (SDR). Ideally, we want to arrive at a population estimation methodology that SDR can implement with their existing work. In this section we consider how best we can include two of SDR’s main streams of work in our project: Admin Based Population Estimates (ABPEs), and IPACS (Integrated Population and Characteristics Survey).

Including ABPEs in the simulation

In simV2, we simulate a “combined admin data source” – where, being on any of the single admin sources means you are included in the combined one. In combination with the assumption of perfect linkage and no misclassification errors, this makes a very high-quality combined source. As we have mentioned, this is highly unrealistic, and does not adequately reflect the ABPEs that SDR produce.

In order to reflect the ABPEs better, we believe we should:

* Introduce more errors that affect the combined admin source (e.g. data linkage errors)
* Begin to model the processes by which SDR create their ABPEs (i.e. exclusions)
* Make sure admin data patterns obtained from the simulation study reflect the patterns seen in ABPEs (benchmarking)

We also think it is worthwhile changing how we have been conceptualising the estimation process. Rather than looking for one best process for making ABPEs, to be coupled with a single best estimator, we would like to use the simulation to create a family of ABPEs that demonstrate a range of combination approaches, and couple these with a variety of estimators. Our intuition is that there will be particular ways of preparing the data that suit particular estimators – for example, if we were able to trim over-coverage out of a combined admin data source, we could apply DSE.

Similarly, we could expand this strategy to include more than one flavour of IPACS. For example, if we were able to run a survey to estimate lambda for IPACS, we could use this information to run Patrick Graham’s method, allowing us to tolerate over-coverage in an ABPE.

How to include IPACS

At present, we do not expect to have much influence over how IPACS is designed or carried out. However, it is clear that this survey is intended to be one that is collected over time, in a wave-form.

Given this, we would like to develop this longitudinal collection in the simulation, and begin searching for estimators that can deal with this type of sampling. Furthermore, since admin data is gathered over time, rather than at a single point, there is a strong motivation to develop this side of the project.

Despite our limited ability to determine what IPACS will entail, we note that this survey is being proposed as a necessary validating source of data to enable admin data estimation. One way in which it could prove helpful is to supply information in some of the areas where we have no corroborating data source – for example, to reconcile household and address (as mentioned in Section 4. Development: Making improvements, and the next version).

Apart from relating our work to ABPEs and IPACS, we would like to note that the simulation/estimation work could potentially be beneficial to SDR and ONS in other ways. For example:

* The simulation is already taking into account both stocks and flows – this may be of use in the hybridisation work that SDR intend to carry out at a later date
* Recently, work in Methodology has been developed for an admin data error framework[[10]](#footnote-10), capturing the ways in which errors can collect in admin data; once generalised, this work could potentially be brought to bear on the simulation, if we were to continue expanding the design
* A new project into using a rolling model and fractional counting to arrive at admin data estimates is being developed in Methodology – this could be tested as part of our range of estimators, if we are able to develop a simulation that is sufficiently realistic

**Section 7. Developing, longer term**

In this concluding section we consider the overall trajectory of this work. In particular, we focus on how our thinking has developed since beginning this project, and how we expect this to shape our research path.

The task at hand is complicated

The task of creating population estimates using admin data is difficult. We do not think this is surprising, since the admin data available is idiosyncratic of England and Wales, and of the time during which it was created; similarly, the population we wish to estimate is unique to its time and place. As such, the population is a moving target, and admin data is an inconsistent capture device. Thus, the task of estimation becomes a combination of both the underlying truth and the extent to which one can “see” it with the data, and cannot be easily boiled down to a single one-size-fits-all method.

How to build an estimation strategy – changes in thinking

Given the complexity of our task, we are beginning to think about an admin data methodology as involving a balance of parts.

Different estimators can be tailored to a certain extent, but – ultimately – we expect to use a combination of them, balancing their strengths and weaknesses (e.g. especially robust to over-coverage that cannot be weeded out through exclusions). This reflects our proposed approach towards incorporating ABPEs in our work: rather than concentrating on a single best ABPE, developing a family of them for use with specific estimation strategies.

Similarly, we expect that the admin data available to us will not be comprehensive and complete, but will require careful work to understand how we can extract and integrate the information that they contain.

We need to aim towards a realistic simulation

At the outset of this project, we envisaged the simulation as being a way to compare estimators. We now understand that in order to do this, the simulation needs to re-create the problems that arise in admin data due to how it is gathered (e.g. over and under coverage). Furthermore, whilst it is useful to include these problems in a generic sense, it is important that we reproduce the problems that afflict England and Wales, specifically.

We hope to achieve this through a combination of data and design: in the design, we aim to recreate the main mechanisms that lie behind the creation of admin data; and we inform these mechanisms using appropriate data.

We have already discussed that our ability to test estimators is only as good as the capability of our simulation to create a sophisticated and realistic estimation problem. Without a sufficiently advanced simulation, it will be difficult to make the case for one estimator or another conferring specific advantages when estimating from admin data.

But how far we can make the simulation “realistic”?

In order to do this, we would need to adequately capture the various mechanisms by which admin data is produced in our design, and find suitable supporting information and data to elucidate these mechanisms. It is not clear that either of these goals are feasible.

Making best use of admin data

We expect that to make best use of admin data, we will need to consider it on its own terms. That is, an underpinning research programme into admin data is vital, to find patterns, to describe how it changes over time, and to spot the mechanisms that indicate how it has been gathered.

However, no matter how well we understand our admin data, we suspect that it will not be enough on its own, and that we will need to incorporate data that is qualitatively different. For example:

- survey data could be used to validate assumptions about admin data, and possibly help to estimate characteristics that are not generally gathered by admin data

- data that is generated in a more automatic way, such that capture probabilities and error are less subject to human action

In conclusion, this project represents a small step on a long research path. However, we believe that our current research strategy of balancing design and data in an iterative process will help us to progress in a logical and consistent way. We have a wide range of questions to answer, but we expect these to become tractable as underpinning research allows us to understand our admin data better.

**Appendix: 1. Data preparation details**

In preparation for the simulation, each data source undergoes some cleaning or alteration. Below are further details about the data used, and about how it was prepared.

Census 2001 for base population:

* An incomplete set of variables are available in this data source; from these, we use variables for sex, single year of age, household ID, OA, and LAD. Household ID is derived from a Census variable “form\_ID”.
* Age is capped at 90, because this is the maximum age in the admin data counts; all of those older than 90 are given a new age of 90.
* Some records are excluded:
	+ Pupils not at their term-time address (tt\_address\_ind!='2')
	+ Communal establishments (person\_type=='P', Enum\_D!='WW01')

Migration data:

* International data is for the period ending mid-2016, and is stratified according to direction of flow, sex, and age group (0-15, 16-24, 25-34, 35-44, 45-54, 55-64, 65 and over)
* This data is based on the international passenger survey
* Data for moving people between LAD7 and E&W is taken from national (internal) migration figures that are stratified by flow, sex, LAD, and single year of age.
* This data is based on administrative data, primarily health data, and involves a complex scaling process, which results in the number of moves taking non-integer values.
* We aggregate this data according to flow, sex, and ten-year age group (0 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70, 71 and over). This stratification was chosen to improve household allocation – i.e., we needed strata to be sufficiently broad.

Admin counts data

* Data is for the year ending 2016 – we note that this does not indicate a clear-cut point in time, as there are differences between sources as to:
	+ when, exactly, each source has been extracted from the admin system that gathers it;
	+ the period over which activity rules are applied (e.g. 6m either way, or 12m prior to the data being collected by the system)

**Appendix: 2. Comparator estimators**

For simV2, we are running the simulation with 4 “comparator” estimators. This section provides details about how we calculate estimates for the total LAD7 population with these estimators.

1. Horvitz-Thompson (HT)

For this estimator, we have two forms: one supposes that we have an adjustment factor for household non-response, and one assumes that we do not. The basic form of this estimator is:

$$HT=\sum\_{k}^{}\frac{y\_{k}}{π\_{k}}$$

The survey unit in our simulation is households (k). For our HT estimator, we weight the survey responses ($y\_{k}$) by the inverse of their selection probabilities ($π\_{k}$). In the case of a simple random sample, everyone has the same probability of selection (0.01); in the case of our clustered survey, sampling probabilities are calculated with the aid of the “mstage” R package, for our chosen survey structure.

Our adjustment factor appears as a second weight, equal to the inverse of the known household-nonresponse probability. The weighted responses are summed up within households, and then summed over all households.

1. Ratio Estimator

This estimator has been coded with the following form:

$$Ratio =\frac{\sum\_{k\in s}^{}y\_{k}}{\sum\_{k\in s}^{}x\_{k}}. X$$

For the ratio estimator, we sum responses to the survey ($y\_{k}$), and scale them by the sum of admin counts in surveyed households ($x\_{k}$), where households in the survey are notated as ($k\in s$).This ratio is multiplied by the total number of admin counts ($X$).

1. Dual System Estimation (DSE)

We carry out the simplest version of this estimator:

$$DSE=\frac{y\_{1•}y\_{•1}}{y\_{11}}$$

For DSE, we require that individuals in our simulation have a flag for being in the admin data (in our case, in the “combined admin” source), and a flag for being in the survey. The count of individuals who are captured in the admin data alone ($y\_{1•}$) is multiplied by the count of those who are captured only in the survey ($y\_{•1}$); this is divided by the number of individuals captured in both survey and admin data ($y\_{11}$).

1. Two-stage Estimator

In our current iteration of the simulation, we have coded our two-stage estimator to follow the same logic as used in the estimation approach for the England and Wales 2011 population census[[11]](#footnote-11). However, in our version, we use Horvitz Thompson at the lower level (OA), and use a Ratio estimator to bring these estimates up to the total population level.

**Appendix: 3. Proposed research questions**

|  |  |
| --- | --- |
| **Research topic** | **Use in the simulation** |
| basic statistics for the presence of individuals – their admin data footprint (e.g. demographics) | What sort of admin populations/ presence should the simulation arrive at? (benchmarking – year-specific) |
| descriptions of missingness across demographics/source | What sort of missingness should we simulate in? |
| frequency and regularity of activity events (and maybe some thought about how indicative different admin data events are of actual activity) | Simulating how frequently people use an admin source  |
| presence and activity across sources, and across time –describing **patterns** and how **stable** they are over time | Finding and applying patterns in the data – points of stability (benchmarking/ validating) |
| Deeper research into coverage patterns by linking with other sources – e.g. surveys | Validating our assumptions of what coverage “should look like”, in a year-independent way |
| Across research, finding **proxy variables** for coverage that are described by the admin data (e.g. like mean churn) | Some soft “internal validation” of admin data (by source)  |
| **Patterns of dependence**: for individuals between sources; individuals within the same household for a source (i.e. patterns in admin data footprint shape) | Build dependence patterns into how individuals are allocated outcomes for admin data – e.g.  |
| **Household** stability and mobility (n.b. as it truly is, and how this is reflected in the admin data) | To inform how people move, which affects OC/UC – see Next Version |

**References:**

Abbott, O (2009). Population Trends, 137, pp25 - 32. 2011 UK Census Coverage Assessment and Adjustment Methodology.

Graham, P; Lin, A. (2017) Small domain population estimation based on an administrative list subject to under and over-coverage. Published for ISI, Marrakech. Available on request.

Graham, P; Lin, A. (2018). Bayesian and approximate Bayesian methods for small domain population estimation from an administrative list subject to under and over-coverage. Statistics New Zealand. Unpublished Internal Report. Available on request.

Rosenthal, J.S (2000). Far East Journal of Theoretical Statistics 4, pp207-236. Parallel computing and Monte Carlo algorithms.

Lambert, P et al (2005). Statistics in Medicine, vol 24, pp2401–2428. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS.

1. https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme [↑](#footnote-ref-1)
2. https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject [↑](#footnote-ref-2)
3. Formerly known as Statistical Population Datasets (SPDs) [↑](#footnote-ref-3)
4. PAYE: Pay As You Earn, HESA: Higher Education Statistics Agency (students in tertiary education), PR: Patient Register (GP records) [↑](#footnote-ref-4)
5. Graham, P; Lin, A. (2017) Small domain population estimation based on an administrative list subject to under and over-coverage. Published for ISI, Marrakech. Available on request. [↑](#footnote-ref-5)
6. Graham, P; Lin, A. (2018). Bayesian and approximate Bayesian methods for small domain population estimation from an administrative list subject to under and over-coverage. Statistics New Zealand. Unpublished Internal Report. Available on request. [↑](#footnote-ref-6)
7. Rosenthal, J.S (2000). Far East Journal of Theoretical Statistics 4, pp207-236. Parallel computing and Monte Carlo algorithms. [↑](#footnote-ref-7)
8. Lambert, P et al (2005). Statistics in Medicine, vol 24, pp2401–2428. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. [↑](#footnote-ref-8)
9. [https:/www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/datasets/internalmigrationbyoriginanddestinationlocalauthoritiessexandsingleyearofagedetailedestimatesdataset](https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/datasets/internalmigrationbyoriginanddestinationlocalauthoritiessexandsingleyearofagedetailedestimatesdataset) [↑](#footnote-ref-9)
10. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsworkingpaperseriesno19anerrorframeworkforlongitudinaladministrativesourcesitsuseforunderstandingthestatisticalpropertiesofdataforinternationalmigration [↑](#footnote-ref-10)
11. Abbott, O (2009). Population Trends, 137, pp25 - 32. 2011 UK Census Coverage Assessment and Adjustment Methodology. Retrieved from <https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howwetookthe2011census/howweprocessedtheinformation/coverageassessmentandadjustmentprocesses/censuscoverageassessment_tcm77-189757.pdf> [↑](#footnote-ref-11)