MARP January 2021

Population estimation without a Census: update

Rosalind Archer, Rob North, Aidan Metcalfe

Introduction

In this paper we present an update for our project to estimate population size by local authority and age-sex using administrative data and a survey, in the absence of a census. Our aim is to develop an estimation methodology, which contributes to the National Statistician's recommendation in 2023, and the development of Admin Based Population Estimates (ABPEs).

To do this we have built a simulation framework that creates a population, their administrative data, and a survey; this observed data is used to test the behaviour of a range of estimation methods. Currently, the framework is still relatively simple, in comparison to the complexity of admin data. We cannot say at this stage which estimators should be used in an admin estimation methodology; however, given what admin data may be available now and in the future, the framework can tell us how well estimators are likely to perform.

For this reason, a core message of our paper is that the simulation framework cannot stand alone. Underpinning research is required to describe which scenarios are feasible and likely; the simulation framework can then demonstrate which estimation methods are useful in such scenarios, and which scenarios are preferable. This is how the simulation framework will be able to offer relevant guidance in developing the best strategy for estimating population size for local authorities by age and sex.

In the current version we have explored scenarios concerning over-coverage in the admin data, and non-response in a supporting survey. At this stage, it seems most likely that neither over-coverage in the admin data nor non-response in a supporting survey can be entirely removed. We recommend that future research continues to prioritise these types of scenarios and that the next most important steps will involve addressing issues of dependence, and heterogeneity of capture probabilities for both survey and administrative data.

Background

In Estimating Population Size without a Census (EAP 129)[1] we developed a framework that enabled various patterns of under and over coverage in administrative data to be simulated. Alongside this, a survey with a simple pattern of non-response was created to be used in various estimators to estimate population size. We found that well-known estimators performed as expected, and we successfully implemented a Bayesian estimation approach that was developed at Statistics New Zealand.

For this paper, the key further developments in the simulation framework and estimation methods include:

a) Changes in the simulation design:

[1] https://uksa.statisticsauthority.gov.uk/about-the-authority/committees/methodological-assurance-review-panel-census/papers/; see also EAP 130: Estimating population size without a census results supplement

- expansion of the simulation to a longer time period
- adoption of a microsimulation approach
- simulation of population change to include births and deaths
- further mechanisms for generating admin data - "historical admin" and duplication
- further mechanisms to explore individual non-response in the simulated survey – differential non-response, and probability of response being dependent on admin data capture

b) Changes in estimation, to include a logistic DSE approach, and a weighting classes approach

c) Changes in the data used for the starting population (base data):

- 2001 Census data are inflated to 2011 benchmarks for age, sex, and average household size, to stand for the population in 2011
- 5 Local Authority Districts (LADs) were chosen to provide a variety of coverage patterns

d) new scenarios to test our estimators, based on a limited number of coverage patterns, and on a variety of non-response types

We found it necessary to restrict our implementation of this design, for reasons of time and resource. In doing this, our intention was to reduce computation and control variation, whilst still exploring the scenarios that we were most interested in.

Simulation design overview

The design of this improved simulation framework can best be described as a journey, which takes the following course:

- it begins with the **base data**, which is read in and is **inflated** to 2011 benchmarks
- the initialised data is passed into a **cycle** of events that both advances the base population through time ("ageing"), and produces admin data for that population
- after the simulation cycle, at T_i, the admin data from all previous processes is brought together, and rules are applied to create a single "admin list" from the separate sources
- a sample of households is taken from the aged-on base data, and a **survey** is simulated, with non-response, and we may allow **duplication** of the admin over-coverage
- the simulated population and their admin data are then sent to our range of **estimators**
- and, lastly, measures of performance are calculated

This journey is represented schematically in Figure 1.

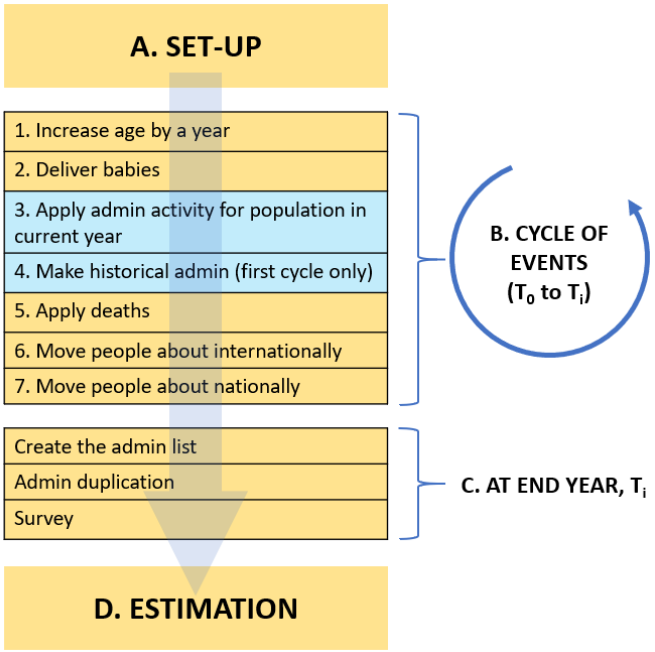

Figure 1: Simulation journey

Base data and inflating

In the first step of the simulation, Census 2001 data for a given LAD were inflated to 2011 benchmarks. This data included variables for LAD, age, sex and a household identifier. Our benchmarks for the number people per age-sex group were derived from Mid-Year Estimates (MYEs) for 2011; and our benchmark for average household size from the 2011 Census . We used 38 (approximately 5-year) age-sex groups, in keeping with a disaggregation common to previous research and the MYEs. In the future we expect to make use of Census 2011 data, and this step will no longer be required.

Simulation cycle

After inflating, the base data is ready to be fed into the main simulation cycle. Every iteration represents a year, during which the population is aged-on and may interact with admin sources (i.e. have "admin activity"). We chose to begin the cycle in mid-2011 and finish in mid-2016; these timepoints were chosen to facilitate alignment with existing data sources (e.g. Census 2011), and previous research into administrative data quality.[2]

As part of ageing the population we simulate births, deaths, and migration. Births and deaths are enabled by using transition probabilities based on published rates, and then aligning to published totals for that LAD/year. Migration is mostly achieved by moving whole households, to meet

[2]https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21

published benchmarks for individual moves by age group and sex. Where an LAD has a significant student population, we move a portion of student-aged individuals singly.

None of these population events have been directly linked to simulated admin activity.

Admin data capture

During each cycle, individuals in the population may be captured by any of five admin sources (HESA, PR, CIS, SC, and births)[3]. Probabilities for whether an individual is captured are derived from 2016 admin data counts for "active" admin records, where the definition of an "active" record is based on activity rules used to make Admin Based Population Estimates[4]. These admin data counts are compared to MYEs for 2016, to arrive at capture probabilities that are aggregated according to age, sex, and LAD. Those who immigrate into the population during a given cycle may not be captured, and contribute to under-coverage in the admin data; however, they may be captured in the simulated admin data in future cycles.

Historical admin data is created during the first cycle only, using the starting population as a donor pool. This process is similar to that described above, but is based on 2016 admin data counts for "inactive" admin records. It is otherwise completely independent of admin activity capture.

At the end of the whole cycle, we apply simple rules to the simulated admin sources to decide whether an individual is captured in an "admin list". Before simulating the survey, we may duplicate data by inflating existing over-coverage.

Survey

We simulate a survey with household as the sampling unit, and a simple random sampling design. We developed 5 non-response options: full response; simple household non-response; household non-response with simple individual non-response; household non-response with differential individual non-response; and household~~s~~ non-response with differential individual non-response, where the individual non-response depends on admin capture.

In our current results we fixed the survey sample size at 2% of all households, household non-response at 50%, and the probability of individual response within responding households is centred at 0.8. These assumptions are based on what might be available from a population coverage survey, and will be revised in future work in line with the most current thinking. -Differential non-response for individuals was achieved by basing response probabilities on sex, and on an artificial non-response variable.

Estimators

After the survey module has been run, we have the data required to use various estimators to estimate population size. Making some further assumptions, in particular around linkage, we obtain estimates based on the observed data (simulated survey and admin data), and calculate measures of performance over the simulation runs. For this work we include amongst our estimators: Horvitz-

[3] School Census (SC), Higher Education Statistics Agency data (HESA), Patient Register (PR), and Client Information System (CIS). The CIS is admin data from the Department of Work and Pensions, and contains individuals who have National Insurance numbers.

[4] ABPE V3 https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21

Thompson, a Ratio estimator, basic Dual System Estimator (DSE), a logistic DSE approach, two weighting class estimators, and two Bayesian estimators based on work from Statistics New Zealand[5,6].

The logistic DSE approach is based on a method used for the US Census[7], which uses three logistic models to predict the probability that an individual is captured in the survey, in the admin, or in both. Our first weighting classes estimator uses admin data to estimate household non-response, and assumes household linkage between survey and admin; a second includes a further adjustment for individual non-response, and assumes individual-level linkage.

For estimates stratified by sex only, the logistic regressions for our logistic DSE estimator were specified with sex as a covariate, and for the scenario with differential individual non-response, we also included our non-response variable as a covariate. For estimates stratified by age group and sex, we followed the same specification for the high-level results, with the addition of age group.

Seeding

We implemented seeding for all runs, such that every ith run per scenario involves the same survey sample, and responding populations.

Adjusting for household non-response

We calculated a weight to allow for household non-response, which is the inverse of the proportion of responding households. We allowed the use of this weight because we believe that it is feasible to know how many sampling units respond in a survey. Under current conditions, where household non-response is random and household is aligned with address (the usual sampling unit for surveys), we expect this weight to work very well. We do not adjust for individual non-response, as we do not think that this information would be easily available. The household non-response weight is applied to HT and Bayesian estimators.

Some key assumptions and simplifications

While some assumptions and simplifications are necessary to any simulation, here we describe those that are most significant, and will have an impact on either inflating or deflating the performance of the estimators. A fuller discussion is provided in the Appendix (in particular, Annex 2).

The design for simulating admin data is overly simple: each individual is only allowed one admin activity event per year, for each source; and the probability of an individual being captured by each admin source is stratified by only age, sex, and LAD. Also, admin captures are assumed to be completely independent events: across individuals, and for an individual – across sources, and across time. These assumptions will result in some of the estimators understating both bias and variance.

The data being used to simulate almost all admin data are based on just one coverage pattern in one year, which is known to be imperfect. Furthermore, the notion of using coverage patterns to develop

[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.
[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.
[7] https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g10.pdf

an admin data methodology is problematic, as we have no clear evidence that past coverage patterns are predictive of future ones.

We assume independent captures for individuals across admin data and survey. We also assume that both admin and survey data have no errors – in particular, this assumes perfect linkage at any level. This is problematic, as we know perfect linkage is unlikely, and it is an important assumption for almost all estimators. Independence assumptions are key for unbiased capture-recapture type estimators such as DSE.

We generally draw on the existing population as an appropriate donor pool, which may not always be appropriate (e.g. for in-migrating households).

We simulate very little in terms of household dynamics, with changes occurring through births, deaths, and (sometimes) when student-aged individuals leave existing households. Also, we effectively conflate the concepts of: household, as defined in Census data; sampling unit (address); and UPRN, which is a variable commonly available in admin data, roughly aligned with address.

Current Implementation

Admin data capture is a complex process, and so requires a relatively complex simulation if we want to explore the mechanisms behind it and their effect on estimation. However, complexity can make it hard to disentangle the effect of each mechanism, and increases the number of ways that the simulation can be run, as well as the time required to run end-to-end. It also makes the task of manipulating the simulation more difficult.

Therefore, we chose a selection of scenarios in an attempt to balance these concerns, and we prioritised those scenarios that we believed would best test our estimators. In particular, these involve different combinations of over-coverage and non-response. Table 3 shows where we chose to restrict the simulation to create our existing results. All events leading to the survey – the generation of the population and admin data – were run just once.

Table 3: The parts of the simulation that were allowed to vary for the current results

Step	**Description of variation in simulation**	**Is this step iterated over in the current results?**
Inflating the base	The selection of which households to duplicate	No
Deliver babies	Selection of mothers	No
Admin activity	Whether individuals in the population interact with admin sources	No
Deaths	Selecting people for deaths	No
International migration	Selecting households and individuals to move	No
National migration	Selecting households and individuals to move	No
Admin duplication	Selecting admin records to duplicate as overcoverage	Replaced by set coverage patterns – used to adjust data
Survey	Selection of households for survey (sampling) Selection of responding households Selection of responding individuals	Yes

Before the simulated data were passed to the survey we adjusted them to meet LAD-level benchmarks for over-coverage, under-coverage, and net-coverage. In order to test a variety of observed coverage patterns, we selected five areas to explore: Cambridge, Manchester,

Northumberland, Peterborough, and Westminster. The process of adjusting the data to coverage benchmarks, and our choice of areas, is described more fully in the Appendix (Annex 3).

Lastly, in order to allow us to pick apart the effects of non-response and over-coverage, we allowed an option for over-coverage in our adjusted data to be wiped out (i.e. simply removed). This also affects the total number of simulated admin data records.

By controlling variation in this way, we arrived at the scenarios listed in Table 4. Our hypotheses for how our estimators would respond to these conditions are included in the Appendix (Annex 4).

Table 4: summary of scenarios - combinations of administrative data over-coverage and survey non-response

Scenario	Over-coverage	Household non-response	Individual non-response
1	adjusted coverage pattern	none	none
2	adjusted coverage pattern	50%, flat	none
3	adjusted coverage pattern	50%, flat	20%, flat
4	adjusted coverage pattern	50%, flat	20%, differential
5	OC wiped out - nil	none	none
6	OC wiped out - nil	50%, flat	none
7	OC wiped out - nil	50%, flat	20%, flat
8	OC wiped out - nil	50%, flat	20%, differential

A key point is that the full machinery of the simulation framework is not yet being fully leveraged. This means that we can support more simulation questions than we pose in this paper, which is why we would recommend a further implementation period.

Further details can be found in the Appendix on the simulation design (Annex 1) and on its current implementation (Annex 3).

Results

The results we present here are for males in the Cambridge local authority, which has a true population of 63,701, for 100 runs.

Tables 6a-6h show bias, Relative Squared Error, and Relative Root Mean Squared Error, for men across all scenarios; similar tables for women can be found in the Appendix. Total estimates for the population are not currently available, but will be included in future iterations. Further results, showing estimates and measures of performance over ten-year age group and sex can also be found in the Appendix (Annex 5).

Measures of Performance

Measures of performance 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.

As in previous research, we recognise that we are comparing estimators of quite different types – design-based, Frequentist, and Bayesian. At this stage of work, where none of our estimators have been developed very deeply, we are content to continuing comparing them using these metrics. It will be a future challenge to maintain a level playing field for these estimators as this work progresses, as we better develop current estimators or include new approaches.

Table 6a. Estimator performance for Cambridge, males, Scenario 1 (admin undercoverage = 13%, overcoverage= 16%, survey size= 2%, perfect survey response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	Relative Root Mean Squared Error (RRMSE)
HT full response	**63701**	**63544**	**2.79**	**-0.25**	**2.8**
Ratio full response	63701	72351	1.52	13.58	12.05
DSE full response	63701	76224	1.32	19.66	16.48
Weighted class 1 full response	**63701**	**63544**	**2.79**	**-0.25**	**2.8**
Weighted class 2 full response	**63701**	**66947**	**2.77**	**5.09**	**5.58**
Back calculation full response	**63701**	**62979**	**1.88**	**-1.13**	**2.2**
Gibbs full response	**63701**	**63572**	**2.78**	**-0.2**	**2.79**

Table 6b. Estimator performance for Cambridge, males, Scenario 2 (admin undercoverage = 13%, overcoverage= 16%, survey size= 2%, 50% household non response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	Relative Root Mean Squared Error (RRMSE)
HT HH non response	**63701**	**63438**	**3.9**	**-0.41**	**3.92**
Ratio HH non response	63701	72237	2.19	13.4	12.02
DSE HH non response	63701	76121	1.96	19.5	16.43
Weighted class 1 HH non response	**63701**	**63444**	**3.25**	**-0.4**	**3.27**
Weighted class 2 HH non response	**63701**	**66858**	**3.23**	**4.96**	**5.72**
Back calculation HH non response	**63701**	**62906**	**2.75**	**-1.25**	**3.03**
Gibbs HH non response	**63701**	**63487**	**3.88**	**-0.34**	**3.9**

Table 6c. Estimator performance for Cambridge, males, Scenario 3 (admin undercoverage = 13%, overcoverage= 16%, survey size= 2%, 50% household non response, 20% individual flat non-response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	RRMSE
HT indiv non response flat	63701	50801	3.41	-20.25	25.62
Ratio indiv non response flat	63701	57869	3	-9.15	10.52
DSE indiv non response flat	63701	76121	2.16	19.5	16.46
Weighted class 1 indiv non response flat	63701	50824	3.79	-20.21	25.62
Weighted class 2 indiv non response flat	**63701**	**66857**	**3.33**	**4.95**	**5.78**
Back calculation indiv non response flat	63701	50435	1.88	-20.83	26.37
Gibbs indiv non response flat	63701	50868	3.41	-20.15	25.46

Table 6d. Estimator performance for Cambridge, males, Scenario 4 (admin undercoverage = 13%, overcoverage= 16%, survey size= 2%, 50% household non response, 20% individual differential non-response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	RRMSE
HT indiv non response differential	63701	51578	3.5	-19.03	23.76
Ratio indiv non response differential	63701	58750	2.83	-7.77	8.89
DSE indiv non response differential	63701	76191	2.08	19.61	16.52
DSE logistic regression indiv non response differential	63701	76208	2.07	19.63	16.54
Weighted class 1 indiv non response differential	63701	51598	3.7	-19	23.75
Weighted class 2 indiv non response differential	**63701**	**66921**	**3.39**	**5.05**	**5.88**
Back calculation indiv non response differential	63701	50440	1.89	-20.82	26.36
Gibbs indiv non response differential	63701	51647	3.49	-18.92	23.6

Table 6e. Estimator performance for Cambridge, males, Scenario 5
(admin undercoverage = 13%, overcoverage= 0%, survey size= 2%, perfect survey response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	Relative Root Mean Squared Error (RRMSE)
HT full response	63701	63544	2.79	-0.25	2.8
Ratio full response	63701	63805	1.32	0.16	1.33
DSE full response	63701	63805	1.32	0.16	1.33
Weighted class 1 full response	63701	63544	2.79	-0.25	2.8
Weighted class 2 full response	63701	63544	2.79	-0.25	2.8
Back calculation full response	63701	62575	1.35	-1.77	2.25
Gibbs full response	63701	62253	1.76	-2.27	2.92

Table 6f. Estimator performance for Cambridge, males, Scenario 6
(admin undercoverage = 13%, overcoverage= 0%, survey size= 2%, 50% household non response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	Relative Root Mean Squared Error (RRMSE)
HT HH non response	63701	63438	3.9	-0.41	3.92
Ratio HH non response	63701	63719	1.96	0.03	1.96
DSE HH non response	63701	63719	1.96	0.03	1.96
Weighted class 1 HH non response	63701	63460	3.25	-0.38	3.28
Weighted class 2 HH non response	63701	63460	3.25	-0.38	3.28
Back calculation HH non response	63701	62114	1.86	-2.49	3.16
Gibbs HH non response	63701	61635	2.5	-3.24	4.18

Table 6g. Estimator performance for Cambridge, males, Scenario 7 (admin undercoverage = 13%, overcoverage= 0%, survey size= 2%, 50% household non response, 20% individual flat non-response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	RRMSE
HT indiv non response flat	63701	50801	3.41	-20.25	25.62
Ratio indiv non response flat	63701	51045	2.79	-19.87	24.95
DSE indiv non response flat	**63701**	**63719**	**2.16**	**-0.03**	**2.16**
Weighted class 1 indiv non response flat	63701	50837	3.79	-20.2	25.59
Weighted class 2 indiv non response flat	**63701**	**63459**	**3.38**	**-0.38**	**3.4**
Back calculation indiv non response flat	63701	50865	1.88	-20.15	25.31
Gibbs indiv non response flat	63701	50880	3.41	-20.13	25.43

Table 6h. Estimator performance for Cambridge, males, Scenario 8 (admin undercoverage = 13%, overcoverage= 0%, survey size= 2%, 50% household non response, 20% individual differential non-response).

Estimator	Truth	Estimate (average over 100 runs)	Relative standard error	Relative bias	RRMSE
HT indiv non response differential	63701	51578	3.5	-19.03	23.76
Ratio indiv non response differential	63701	51821	2.58	-18.65	23.07
DSE indiv non response differential	**63701**	**63777**	**2.08**	**0.12**	**2.08**
DSE logistic regression indiv non response differential	**63701**	**63792**	**2.07**	**0.14**	**2.08**
Weighted class 1 indiv non response differential	63701	51611	3.74	-18.98	23.72
Weighted class 2 indiv non response differential	**63701**	**63520**	**3.41**	**-0.28**	**3.43**
Back calculation indiv non response differential	63701	50865	1.88	-20.15	25.31
Gibbs indiv non response differential	63701	51658	3.49	-18.91	23.57

RRMSE = Relative Root Mean Squared Error = Coefficient of variation. Coefficients of variation less than 10% highlighted in bold

Formatted: Font: 8 pt, Not Bold

Findings

1. HT: struggles under conditions of non-response

- when the non-response is only at the household level, it is unbiased but more variable than when response is full (recall that for HT and Bayesian estimators we use a household non-response adjustment weight)
- when non-response includes individual non-response, HT is biased negatively as expected because the estimator assumes that the survey count is perfect within a responding household.
- the HT estimator, as expected, has a relatively high variance (RSE) across all scenarios as compared to the other estimators, reflecting that it does not use the administrative data in any way
- under conditions for full survey response and no over-coverage, HT, WC1, and WC2 are equivalent

2. Ratio: sensitive to over-coverage and non-response

- when over-coverage is not present, individual non-response caused estimates to be negatively biased
- when non-response is not present, over-coverage causes positive bias (this can be seen best in the estimation classes with the highest over-coverage - age groups 21-30 and 31-40)
- when both non-response and over-coverage are present, the effects of over-coverage and non-response compensate one another but bias is still present
- under conditions of full response and no over-coverage (or only household non response) our version of Ratio is equivalent to basic DSE

3. Basic DSE

- when survey non-response is present and no over-coverage, DSE works well, as expected
- when administrative over-coverage is introduced, it is positively biased as expected

4. Logistic DSE

- similar to basic DSE, this estimator does well in conditions of non-response, but struggles with over-coverage
- our version of logistic DSE does not particularly outperform basic DSE when comparing between scenarios of flat individual non-response and differential non-response

We had expected to observe that logistic DSE would demonstrate an improvement over basic DSE, under conditions of differential individual non-response. However, as we have not introduced significant differential non-response into our simulation and we are not estimating simultaneously across a number of local authorities this is as expected. Future simulation extensions will begin to show the advantages of this estimator.

5. Weighting classes (1 and 2)

- WC1 does well under conditions of household non-response, but becomes negatively biased when individual non-response is introduced, much like the ratio estimator
- WC2 does well under both non-response conditions
- WC1 remains robust under the inclusion of over-coverage provided there is no or minimal individual non response

When both over-coverage and survey non-response are present, WC2 outperforms DSE. However, this must be taken in the context of this weighting class estimator being comprised of a DSE component for the within-household non-response adjustment as well as the adjustment for household non-response. The latter adjustment is not present in the DSE, and results in this weighting class estimator being less susceptible to the effects of over-coverage in the admin data.

6. Bayesian estimators (Gibbs and Back calculation)

These estimators have not been optimised, but we still offer the following findings:
- Both are affected by non-response, leading to negative bias under conditions of individual non-response (recall that we allow a household non-response adjustment);
- Both estimators are able to adjust well for over-coverage
- Under both over-coverage and individual non-response, these estimators are consistently negatively biased
- under conditions of non-response, the Gibbs estimator is more variable than the back-calculation estimator. We think this is because the Gibbs approach models how the admin data are generated, based on the distribution of characteristics in the observed data, whilst the back calculation does not. We think that the observed difference in results stems from this model not being well specified for our simulated data.

General points

All estimators do well when there is full response and no over-coverage, and none do very well when there is a notable amount of each. This raises a few questions:

- Are any estimators capable of dealing with both over-coverage and non-response?
- How much over-coverage or non-response is "too much"?

The second question becomes particularly important when we consider a wider estimation strategy that could involve a stage to prepare the admin data, or possibly to use the survey in a new way (e.g. to estimate over-coverage). If an estimation strategy involved removing over-coverage, or avoiding non-response, it might be feasible to make use of the estimators we have been exploring; however, it seems unlikely that either problem could be completely eliminated. Under such conditions, two questions become crucial:

1. how much over-coverage and non-response can estimators tolerate?
2. to what extent can we realistically reduce over-coverage and non-response, in practice?

Another point that should be considered is that under conditions of non-response, an indirect effect arises – the number of survey responses is reduced, leading to a smaller number of survey counts and an increase in variation for all estimators. This can be explored in future work by simulating a number of local authorities simultaneously.

Discussion and development

Our main finding is that none of our estimation approaches can deal very well with both over-coverage and within household survey non-response. This means that the following questions are particularly pressing:

- Tolerance of over-coverage and non-response – how much can estimators tolerate?
- How much over-coverage can practically be removed, and how high a survey response can we practically expect, from the observed data?

Both must be answered if we want to provide evidence for developing an administrative data population size estimation methodology. The simulation framework can be developed to answer the first question, however, we anticipate that the future success of this project will require both. We also expect that our future direction can be prioritised into the following steps – exploring how estimates are affected by:

1. Over-coverage in administrative data
2. Survey non-response
3. non-independence: between admin sources, and between admin and survey
4. Homogeneity of capture, particularly in admin data (and so addressing the question of geographical heterogeneity in admin data)

These steps capture our existing findings as well as concerns mentioned above, for example: the insufficiency of a single coverage pattern to simulate admin data, and the assumptions of perfect linkage and independent captures for both admin and survey. They also indicate further important questions to be addressed as underpinning research:

- Are coverage patterns stable over time?
- What is the joint distribution of admin data and survey captures?
- What are the dependency patterns for individual admin capture between sources and over time?

By bringing together the simulation and the right underpinning research, we will be able to use this project to provide direction for our admin data estimation methodology. The simulation allows us to explore estimator behaviour under specific conditions; the underpinning research specifies what those conditions might feasibly be.

Some recommendations for developing the simulation framework:

1. Time and resource did not allow us to fully explore the simulation, but the existing design would support:

- exploration of scenarios where survey capture is dependent on admin capture – initial findings suggest that this can have a large effect on bias and variance, and that the Bayesian back-calculation approach may be more robust to these effects than our other estimators
- inclusion of longitudinal elements – e.g. wave-form survey, longitudinal patterns in admin data capture
- moving away from point-in-time survey and estimation – this is particularly important, as we expect that in reality the survey will be ongoing, and estimates will be made across time. This will also involve considering an additional method for combining estimates across time, using time series methods.

2. Use more tools from the field of microsimulation to support this work:

- develop an emulator (a statistical model to map inputs to outputs) to explore and characterise simulation variation
- sensitivity analysis of parameters
- develop a more standardised description of the simulation (e.g. provenance modelling, or ODD+ (Overview, Design concepts and Details) procedures [8]), to facilitate clearer development choices and communication

3. move to Data Access Platform (DAP):

- computational gains through distributed computing – to facilitate a less restricted approach to using the simulation, and allow us to run more sophisticated estimators
- access to better data – record-level admin data, Census 2011

We recognise that there are some important questions that remain unanswered, but that we expect to address in due course. For example: how shall we scale the work up to create national-level estimates? What other types of estimation approach might be developed? And, how will this work be integrated with sister projects inside ONS?

In particular, there are opportunities to collaborate with our colleagues in:

- Developing how ABPEs are made (Long-term ABPEs, Population Migration Statistics Transformation division (PMST))
- Understanding the errors in admin data (Methodology, Admin Error Framework[9])
- Estimating migration with admin data (PMST, Methodology)
- Rolling estimation methods, such as fractional counting (Methodology)

Going forward, it will become particularly important that we share findings, and that we seek to build up a common understanding of admin data and how it can be used. In doing this we believe it will be possible to build a coherent admin data estimation methodology, which is capable of supporting a transformed statistical system.

[8] Reinhardt, Oliver & Ruscheinski, Andreas & Uhrmacher, Adelinde. (2018). ODD+P: COMPLEMENTING THE ODD PROTOCOL WITH PROVENANCE INFORMATION. 727-738. 10.1109/WSC.2018.8632481.

[9] https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/longitudinal-linkage-of-administrative-data-design-principles-and-the-total-error-framework

References

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.

Office for National Statistics (2015). Developing a weighting-class approach for the 2021 Census

Office for National Statistics (2019). Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016

Office for National Statistics (2020a). Longitudinal linkage of administrative data; design principles and the total error framework

Office for National Statistics (2020b). Births in England and Wales: summary tables

Office for National Statistics (2020c). National life tables: England and Wales

Office for National Statistics (2020d). Deaths registered by area of usual residence, UK

Office for National Statistics (2020e). Internal migration: by local authority and region, five-year age group and sex

Reinhardt, Oliver & Ruscheinski, Andreas & Uhrmacher, Adelinde. (2018). ODD+P: COMPLEMENTING THE ODD PROTOCOL WITH PROVENANCE INFORMATION. 727-738. 10.1109/WSC.2018.8632481.

United States Census Bureau (2012). Census Coverage Measurement Estimation Report: Aspects of modeling

Zhang, L-C. (2019). A Note on Dual System Population Size Estimator