Methods for 2021 Census playbook

Owen Abbott, Methodology Division, ONS.

Cal Ghee, Census Statistical Design Division, ONS.

Contents

Key Messages of Paper

Purpose

- This paper proposes some methods to be used as part of the 2021 Census playbook, which is essentially where the standard design is modified in scenarios where the assumptions underpinning the standard design mean the quality of census outputs is at risk.

Recommendation

- The paper recommends that these methods be used should they be required, although additional work is required to fully specify them if the scenarios become a reality.

Key Asks of CRAG/MARP

- The panel is asked to:
    - Note the context for the playbook contained in the Annexes.
    - Provide feedback on the methods proposed.

0. Executive Summary

The 2021 Census statistical design includes methods which ensure it is robust against minor violations of assumptions which underpin it. However, there are scenarios where those assumptions are strongly challenged, for example extreme coverage failure in the Census for a localized area. In these situations, an alternative methodology is required to produce estimates which may be better than the standard design. This paper outlines the scenarios, and then some methods which may be candidates to use in such scenarios. However, this does not mean that the resulting estimates would meet the overarching census quality criteria should they be implemented – rather they provide a better chance of providing more plausible estimates which may be able to be used as part of the census outputs. The context of developing the 'playbook' is given, showing how they will be governed and how decisions will be taken. The paper also includes information on the main source of alternative administrative information which could be used in the playbook methods.

1. Introduction

The 2021 Census will be subject to both under-coverage (also known as under-enumeration or non-response) and over-coverage. Both, if untreated, will result in biased census statistics. For example, under-coverage will result in a census estimate that is too low.

The standard statistical and operational design of the 2021 Census attempts to reduce both types of coverage error through most of its products, services, and processes, as outlined in ONS (2020). The overarching quality goal is to achieve coverage of 94%, which means 6% under-coverage in the census collection operation. There

are also targets to ensure no local authority achieves less than 80%, and there is minimal variation across and within LAs. The Census Coverage Survey (CCS) is designed to specifically measure the coverage achieved in the census, as described by Castaldo (2018). The CCS is used in the coverage assessment and adjustment process which aims to produce approximately unbiased population size estimates (Racinskij, 2018) and provide an individual level output database which has been adjusted for under-coverage (Whitworth *et al*, 2018).

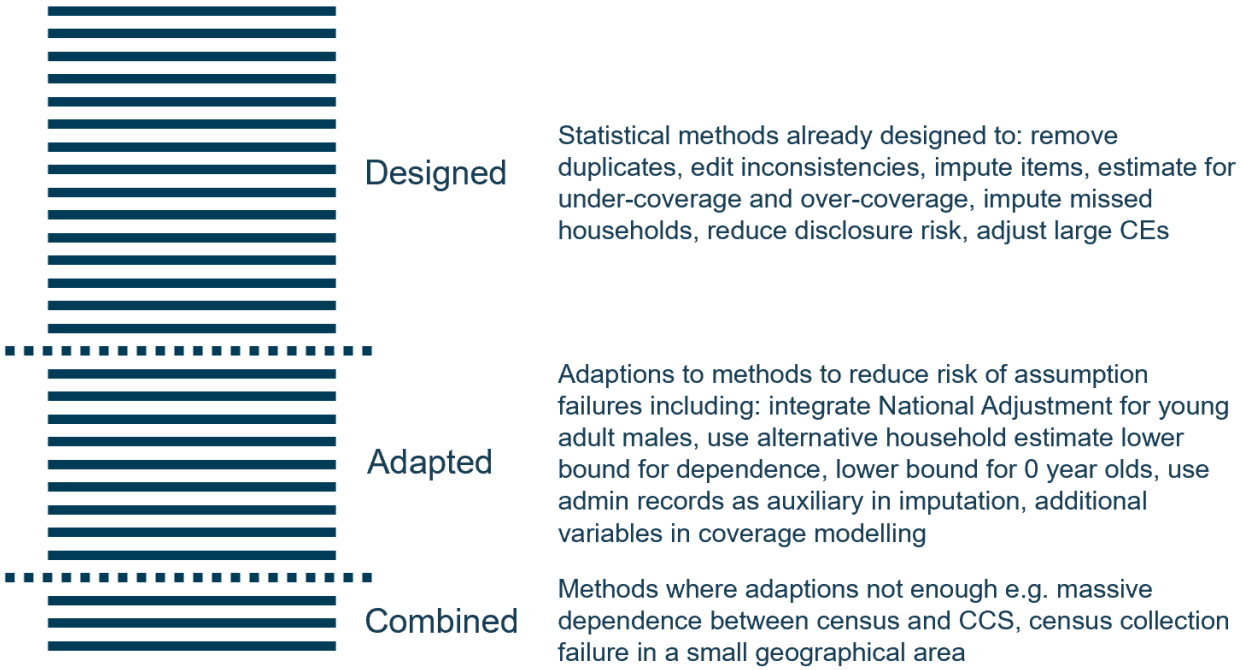
The COVID-19 pandemic has placed additional pressures on the ability of the census to produce robust statistics, due to the increased uncertainty of the population behaviour and risks of lower response or the potential inability to undertake or complete fieldwork in either the Census or CCS. For example, community engagement has not taken place as intended. There are some positives, in that on-line usage has increased and digital networks have developed to support communities during this period.

In addition, the census design does make some operational and statistical assumptions. When these assumptions are not met, the impact can affect the measurement of coverage and thus the population estimates. The impact of this is that the population size could be underestimated (or overestimated). For example, if the census and CCS do not use independent address lists then there is a substantial risk that the CCS will not find addresses which were missed in the census leading to a failure of the assumption of independence required in the estimation methodology, and this leads to an underestimate of the under-coverage.

For many of these assumptions, the standard statistical and operational design itself should mitigate the risk of assumption failure. Experience from previous censuses has provided evidence where these assumptions are most at risk, for examples see Brown *et al* (2006), Large *et al* (2011) and ONS (2013). Adaptions to the methods have already been built into the design where it is known that certain assumptions are likely to not be met, as shown in Figure 1. For example, it is expected that the Census and CCS will not be fully independent, so a method was developed to measure and adjust for that lack of independence by Brown *et al* (2006). This uses several sources to provide an Alternative Household Estimate (AHE) for CCS sampled areas and makes an adjustment to the population estimates using this additional data.

However, it is recognized that not all assumptions can have treatments put in place as part of the standard design as there are far too many extremes or edge cases. This paper proposes a set of combined methods that can be used within the 2021 Census coverage assessment and adjustment process when the assumptions which underpin its statistical design are violated and the existing design does not provide sufficient protection. This includes methods to be used when the census response is extremely low, or the CCS is not possible in some areas due to a localized COVID-19 lockdown. However, it must be noted that in many scenarios the standard design remains the most robust methodology and will produce the highest quality estimates compared to alternative approaches.

Figure 1 – Overall statistical design concept showing designed, adapted and combined methods



This paper focuses on the methods for the playbook, in particular the risk to the population estimates. Further work is required to consider methods where there is significant item non-response which may need a different approach. At present that is out of scope for this paper.

These methods must be seen in the context of the overall approach to planning for scenarios where our standard methods are not robust. The context and some of the work to date are included in the Annexes:

- Annex A: playbook approach
- Annex B: Possible scenarios
- Annex C: Governance
- Annex D: Summary of from International Peer Review of Statistical Contingencies Governance wargames exercise
- Annex E: Alternative data sources, including ACID

Section 2 of this paper provides a reminder of the methods that are part of the standard design which already provide mitigation against the failure of some of the underpinning assumptions. Section 3 then outlines some of the scenarios where there is particular risk to the census quality. Section 4 proposes a spectrum of methods which are modifications of the standard design, which could be used in particular scenarios. In general, these involve a strengthening of the assumptions that underpin capture-recapture methods, and utilization of additional data from, for example, administrative records. Section 5 then brings sections 3 and 4 together with a summary of which methods may be suitable under the different scenarios.

2. Methods that are part of the design

This section outlines all the designed and built-in methods which mitigate against failures of assumptions which underpin the census statistical design.

The basic coverage estimation strategy uses the CCS as the main tool for assessing coverage in the Census. It is designed to be able to measure under-coverage arising from all three sources: missed addresses, household non-response and within household non-response. The CCS is designed to measure this under-coverage, even if it is extreme for some groups (e.g., young non-white unemployed males in rented households deprived areas might have coverage levels of 50%). The sample size is such that it allows accurate estimation across many sub-groups and geographies. The estimation models which are being used in 2021 allows inclusion of additional variables, which provide additional protection to variations in census coverage patterns. This is particularly important given the 2021 Census is primarily online.

As well as measurement of under-coverage, the CCS also provides data for measuring over-coverage although only for duplication and people counted in the wrong location (Racinskij and Hammond, 2019). False returns are not able to be estimated. The Census design also includes an algorithm for removing duplicate returns from the same location.

The under and over-coverage estimation methods are underpinned by linkage between the Census and CCS. High accuracy is required to avoid any bias, and the strategy is to adopt a conservative automated and structured clerical approach to achieve that quality (Shipsey, 2019). Quality assessment of the linkage quality will be a part of that methodology.

Two of the key working assumptions for the estimation methodology are:

- independence between the Census and CCS
- homogeneity of capture probabilities

Previous experience has shown that these assumptions are not fully met, and result in biases in the population estimates. Intuitively, the CCS will be positively correlated with the Census resulting in a lack of independence and an underestimate. The capture probabilities in either the Census or CCS are unlikely to be fully homogenous within the estimation stratum, even within a modelling approach which allows many stratification effects to be included. The two assumptions are related and cannot be disentangled in terms of their impact on the estimates. However, they are relatively well understood within the literature and within the census conext. Brown *et al* (2006) describe the method that is used to make an adjustment for biases in the estimator, which calibrates the estimates to an external estimate to adjust for the bias. The external estimate used in an Alternative Household Estimate, which is an estimate of the number of occupied households. This adjustment estimates the bias due to these two assumptions and any other residual biases which may exist in estimating household under-coverage.

There is also an assumption of independence within-households. Previous studies have shown this to be not significant, and no adjustments have been made. However, the standard design will use the Census Non-Response link study to assess whether there is any bias and make an adjustment using its data should it be required.

These adjustments for bias use data sources which do not provide a full characteristic breakdown of the biases across sub-groups. Thus, a residual bias adjustment is assessed and made at National level, using sex ratio analysis to detect whether there is a differential bias across the sexes that needs adjusting (Smallwood, 2021).

The coverage strategy also deals with under-coverage within Communal Establishments as described by Fraser and Ghee (2021). For small Communal Establishments with 50 or fewer bedspaces, the CCS provides information

on under-coverage in an equivalent way to households albeit with smaller sample sizes. For larger CEs, administrative data and Census field information will be used to directly assess coverage and adjust if required.

Lastly, sitting over all these designed mitigations is the Census Quality Assurance strategy, which assesses the population estimates for plausibility through analysis and comparison with other source of data such as the demographic mid-year estimates and administrative data sources (Large, 2020). This is designed to detect where the estimates may not be plausible and direct further analysis which may result in revisions to the standard design or indicate where a statistical contingency such as those proposed in this paper should be considered.

3. Scenarios where the under-enumeration design is at risk

This section describes some scenarios where the under-enumeration statistical design will be at risk because the existing methods and mitigations outlined in section 2 will not address them, or the resulting quality may not be as good as an alternative approach. These are not likely to be exhaustive but are based on previous experiences both within the UK and internationally such as the issues New Zealand faced in their recent census. For these, it is difficult to give a precise definition or point at which the scenario implies that a non-standard approach would be better than the designed statistical processes. Determination of that point would require extensive simulation studies for each, and there is not the time or resource to devote to that work.

Scenarios where the standard design is still likely to be the best possible strategy are not included. For example, in a scenario where census response is 80 percent, but the CCS is good quality, the standard strategy will still likely provide lower bias and variance than any alternative use involving administrative data. In the 2001 Census, the standard strategy was applied successfully in areas with coverage of 64% (Kensington and Chelsea). Sub-populations with coverage below 80%, such as young males in London, were estimated successfully in the 2011 Census.

3.1 Large scale Census response low, CCS response low, high dependence, and low quality AHE

If the census coverage is much lower than the 94% target, and the CCS response is much lower than its 90% target then there is a risk that lack of independence becomes much more significant. This could be national or within region or (say) Hard-to-count stratum. The strategy for measuring and adjusting for a lack of independence using the AHE would then become more important, and there is a risk that the assumptions that underpin that methodology also fail. The Alternative Household Estimate relies on information from the census fieldwork (especially dummy questionnaires) so that may also be low quality if the census has been poor.

3.2 Matching errors much higher than expected.

The Census to CCS matching accuracy targets have been set high: 0.1 % false positives and 0.25% false negatives. There is a method in place to estimate the levels of error to provide assurance. If they are much higher than these, this will result in bias in the estimates. The measurement will provide information on the levels, but should they be more than (say) 0.5% false negatives then an (unplanned) adjustment would be required. If this were widespread across the whole CCS, then the increased bias should be detectable in the quality assurance process and the national adjustment.

3.3 Unique small area Census failure

This is a scenario where a small area (MSOA, LSOA, OA) has extremely high under-coverage which is due to a factor not included in the general coverage models (for example a localized COVID lockdown, flooding which resulted in population displacement), and it is not a phenomenon repeated across the UK (and thus the CCS might have sampled such cases). The coverage modelling would underestimate the coverage levels, so an adjustment would be required.

These types of occurrences should be picked up during the census operation as areas with extremely low response, so they should be known about. Dummy form information may help identify them during processing should they be missed.

Areas with large concentrations of hard to count populations should not be mistaken for these cases – there will be areas with young, non-white, renting, deprived populations with coverage rates below (say) 60%. The coverage modelling will still be likely to produce robust estimates in these cases, as it did in 2011.

3.4 Significant false returns

Whilst the design mitigates to a degree against false returns (through authentication), there is a risk that responders include false (or erroneous) responses. For example, their pets. These cannot easily be detected. If there are large numbers of false returns, the estimates will be too high.

3.5 Deliberate non-response from a population sub-group in Census AND CCS

If a population subgroup has extremely low response in both the Census and the CCS, and this is due to deliberate evasion, then the estimates will not sufficiently include that population. This is an extreme form of correlation (or heterogeneity) bias in the DSE. This should not be mistaken for the scenario where people misclassify themselves (but are counted).

3.6 Lack of names and date of birth

If there is a campaign to not provide names and dates of birth, meaning the Census-CCS matching becomes exceedingly difficult, as would any linkage to administrative data sources.

4. Potential Methods for a playbook

These methods are organized by the type of scenario that was described in section 3. A spectrum of methods are available here, which range from using existing Census data through to using no Census data. Which should be used is down to the specific situation, but the principle should be to use the census data unless there is evidence to suggest that better results are available from alternatives.

4.1 Adjusted Dual-System Estimate (DSE Adj)

Much like the over-coverage and national adjustments already part of the standard design, the DSE models can be weighted to take account of any measured biases in the DSE. So, for instance, if matching error or false returns are significant and could be measured, then that measurement can be converted into a weight and used to adjust the DSE based models in the standard design. This is because it is known how that type of error can affect the DSE. The quality of such an adjustment would depend on the accuracy of the measurement (e.g., if it is derived through a sampling process to measure matching error or false returns it will have some uncertainty),

and naturally that would then lead to increased variance (although with decreased bias as intended). Such a method could be applied differentially in different sub-population groups or areas if required.

Advantages – Simple to implement, makes full use of coverage patterns measured by CCS, full adjustment using standard methods, minimal changes required to standard design.

Disadvantages – assumes the bias can be measured with a reasonable degree of accuracy.

4.2 Synthetic estimate using Census (C-Syn)

This method is a natural extension of the way in which the standard census estimates are calculated and can be used if the CCS is unavailable in an area (i.e., local authority). The standard coverage models are fitted at national level to the linked Census and CCS. They provide a set of weights by age-sex, hard-to-count, ethnicity, tenure, and other characteristics. These weights can be applied to the census data for an area (say a local authority) to produce a synthetic population estimate even when the CCS did not take place in that area. This assumes that the census response is like other areas, once all the other variables are considered. Therefore, this would work well in a scenario where (say) there is a localized lockdown in an LA which only affected the CCS. The validity of this assumption could be tested by looking at other indicators, for instance dummy response patterns. The Alternative Household Estimate will also provide a quality check and could be used in the same way as in the standard design to detect any residual bias.

Advantages – Simple to implement, makes full use of coverage patterns measured by CCS, full adjustment using standard methods, minimal changes required to standard design.

Disadvantages – only assumption is that census response is like other areas, once age-sex, ethnicity, tenure, hard to count effect are considered.

4.3 Synthetic estimate using administrative data (A-Syn)

This is an aggregate approach which models the relationship between administrative data (e.g., ACID) and the 2021 Census estimates for areas which have (in general) been through the standard design and passed Quality Assurance. The relationship is then used to estimate the population size for an area (e.g., local authority) using the administrative data for that area. It can be used for small areas such as LSOA or MSOAs, as well as whole local authorities. The models would be limited by the characteristics available on administrative data, for instance age and sex and LSOA.

A simple regression approach, fitted at LSOA level, would achieve this, the model would be something like:

$$Cen_{ahg} = \hat{\beta}_{ah} admin_{ahg} + \varepsilon$$

Where Cen_{ahg} are the census age-sex group estimates by hard to count across LSOAs and $admin_{ahg}$ are the corresponding administrative data counts for the LSOAs. The estimated model coefficients are used to estimate the population in LSOAs within the area of interest (if it is a whole LA then it would be for all the LSOAs in the LA) by applying them to the known administrative data counts for that area. The model could be developed by adding additional non-census covariates (for example the IMD) if it does not fit well or adding a regional effect.

Additional variables would help only if they can explain the variation between census estimates and administrative data counts – but this could be explored using 2011 data. Data from multiple administrative sources could be used if it can be shown to improve model fit.

This approach would provide a set of population estimates by age and sex. Household estimates could be applied using a similar approach, perhaps by using the census address frame information, or by using the Alternative Household Estimate for the LA/area, calibrating as appropriate.

Exploration of the feasibility of a revised method for coverage adjustment would be required. One approach would be to use only these age-sex and household estimates as benchmarks, which would potentially result in biased estimates for other characteristics (e.g., ethnicity) as these would be uncontrolled. Some simulation work would need to explore the resulting quality – this may find that under low coverage scenarios these biases are serious enough that the methodology may not result in statistics that would be suitable for publication. This may lead to considering publishing the weights rather than a fully adjusted database for that area.

Advantages – Does not require any linkage. Do not necessarily need to use administrative data with the same reference period, as only modelling relationships (and assuming the relationship holds). Borrows strength from all census areas which are like the failed area.

Disadvantages – It ignores the census data that has been collected in that area for estimation. It assumes the administrative data does not have any specific features in the area (i.e., we assume it is like most admin data in the same type of area). Adjustment quality would be low.

4.4 Weighting Class Estimator with Admin data and CCS (WC-CCS)

Dawber *et al* (2021) describes and explores a weighting class estimator. It is a form of DSE which reduces the linkage required and can be less susceptible to over-coverage in the second source (in most cases administrative data). However, it has a larger variance (and thus wider confidence intervals). It could, for instance, use the administrative data that has already been linked to the census address frame (ACID).

The method would be to use the existing matched Census and CCS information to firstly measure within household coverage, using a standard DSE which only uses data from households which were in both Census and CCS. This could use the modelling approach we have already developed.

Secondly, household non-response would be estimated using administrative data using a weighting class approach. This would also mean the AHE does not need to be used.

Once the Census and admin list are linked at household (address) level, the weights of the age-sex classes were calculated like so for class g:

$$w_g = \frac{Number\ of\ individuals\ in\ class\ g\ in\ admin\ list}{Number\ of\ individuals\ in\ class\ g\ in\ linked\ list}.$$

Using these weights, the weighting class adjusted estimate of the population size is:

$$\hat{y}_{WC} = w_g \hat{y}_g$$

where $\hat{y}_g = \sum_{i \in g} y_i$ and y_i is 1 or 0 depending on whether the i-th individual responded to the Census or not. The classes can be refined to try to capture patterns of household or person non-response, but this is limited to

variables that are available on both the administrative and census data at household level (for instance hard-to-count, accomodation type).

Pros – Uses the administrative data to support the estimation of census non-response. No further individual level matching is required.

Cons – May still have small biases due to over-coverage and heterogeneity in the weighting classes, as well as bias for households not on the census frame. Will have larger variance (perhaps double that obtained from a DSE) due to use of weighting classes. This option is really only realistic in the situation where the Census response is very poor AND the CCS is also poor.

4.5 Weighting Class Estimator with Admin data and other survey (WC-Admin)

This is as method 4.3, but if the CCS is not available some other data is required to estimate for within-household non-response.

This could be, for instance, the data from the CNRLS which uses social surveys to measure within household coverage. Given the data is linked, the method would use a standard DSE which only uses data from households which were in both Census and Social surveys. This would likely be a simple DSE unless time allowed for development of a modelled approach.

A second option is to use admin data to estimate within household coverage, but this would require an administrative data source with zero over-coverage.

Pros – Uses the administrative data to support the estimation of census non-response. No further individual level matching is required. CNRLS linkage is also already planned.

Cons – May still have small biases due to over-coverage and heterogeneity in the weighting classes. Also makes assumption that the Census and Surveys are independent, which may not be the case if the surveys are mainly online. Larger variance due to use of weighting classes. This option is only realistic in the situation where the Census response is very poor AND the CCS is also poor or not available.

4.6 Dual System Estimation using administrative data (A-DSE)

This is the approach that Statistics New Zealand used in their census. The census data that has been collected is linked at individual level to the administrative data, and a simple stratified DSE is calculated, stratified by area and age-sex. Administrative data with virtually zero over-coverage would have to be used to eliminate over-coverage bias in the DSE.

Coverage adjustment may be possible as per option 4.3.

Pros – Uses the existing census data.

Cons – Need administrative data for the right reference period to be available. Individual level linkage is required. May still be biased, likely heterogeneity biases and any unresolved over-coverage. Cannot stratify beyond variables on administrative data so age, sex, and area.

4.7 Use Admin data directly (A-Direct)

This approach does not use any census data – it simply takes whatever the administrative data (e.g., ABPE) provides on population size estimates, by age and sex. Coverage adjustment may be possible as per option 4.3.

Pros – Simple to do, if a suitable ABPE is available.

Cons – Does not use any census information, so may be difficult to defend unless other methods are not possible. At risk of introducing over-coverage if it is present in the administrative data. Would need to explore uncertainty measures.

5. Scenarios and methods

The following table shows the various scenarios where the standard design is unlikely to fully work, or will struggle to provide the best possible estimates. For each scenario, the potential methods that could be employed to provide an alternative set of estimates are indicated within the columns of the table. The methods are ordered by likely quality levels, as assessed by the authors, from left (highest) to right (lowest).

	Method							
Scenario	**Standard**	**DSE adj**	**C-Syn**	**A-Syn**	**WC-CCS**	**WC-Admin**	**A-DSE**	**A-Direct**
Large scale Census response low, CCS response low, high dependence and low quality AHE	X	X	X	X	Yes	Yes	Yes	Yes
Matching errors higher than expected	X	Yes	Yes	Yes	X	X	X	X
Significant false returns	X	Yes	Yes	Yes	X	X	X	X
Deliberate non-response from a population sub-group in Census and CCS	X	Yes	X	X	X	X	Yes	Yes
CCS Not available nationally	X	X	X	X	X	Yes	Yes	Yes
Unique area level Census response extremely low	X	X	X	Yes	Yes	Yes	Yes	Yes
CCS poor or unavailable in an LA	X	X	Yes	Yes	X	X	X	X
Low coverage of a population subgroup (e.g. students)	X	X	X	Yes	Yes	Yes	Yes	Yes
Lack of names and date of birth	X	X	X	X	Yes	Yes	X	Yes

In all the localized low coverage scenarios, the small area where the census has failed is excluded (i.e., it is put into a separate stratum) from the usual coverage assessment processes. The specific method applied for estimating (and adjusting) is applied only within that stratum.

The methods presented in this paper are not comprehensive but give an indication of the sorts of methods that could be realistically applied in these situations. However, they have not been fully specified and researched in

depth, as the focus has remained on the standard design. Nevertheless, if time permits further work will concentrate on those which may have additional benefits in that they are easy to implement and can work alongside the standard design. For example, the synthetic estimates using administrative data could be applied for every local authority to provide an alternative population size estimate which may be useful in the standard quality assurance process.

Lastly, there is still some work to do to explore the feasibility of the coverage adjustment methodology and whether that can still produce robust results under some of these alternative methods. That research is underway as part of a suite of simulations exploring performance under different coverage levels and numbers of benchmarks.

6. References

Brown, J., Abbott, O. & Diamond, I. (2006) Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society Series A*, 169, 883–902

Castaldo, A. (2018). 2021 Census Coverage Survey Design Strategy. UKSA Census Methodological Assurance Review Panel paper EAP 103. Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP103-Census-Coverage-Survey-Design-Strategy.docx

Dawber, J. (2021) Weighting class adjustment and population size estimation for Census 2021 – a simulation study. UKSA Census Methodological Assurance Review Panel paper EAP 1XX. Forthcoming.

Fraser, O. and Ghee, C. (2021) Design of Address Frame, Collection and Coverage Assessment and Adjustment of Communal Establishments in 2021 Census. UKSA Census Methodological Assurance Review Panel paper EAP 141. Available at https://uksa.statisticsauthority.gov.uk/publication/eap141-design-of-address-frame-collection-and-coverage-assessment-and-adjustment-of-communal-establishments-in-2021-census/

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

Office for National Statistics (2013) Coverage assessment and adjustment evaluation. Office for National Statistics. Available from http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/evaluation---coverage-assessment-and-adjustment.pdf

Office for National Statistics (2020) Statistical Design for 2021 Census in England and Wales. Office for National Statistics. Available at https://www.ons.gov.uk/census/censustransformationprogramme/censusdesign/statisticaldesignforcensus2021englandandwales

Račinskij, V. (2018) Coverage Estimation Strategy for the 2021 Census of England & Wales. UKSA Census Methodological Assurance Review Panel paper EAP 105. Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP105-Coverage-Estimation-Strategy-for-the-2021-Census-of-England-and-Wales.docx

Račinskij, V. and Hammond, C. (2019) Overcoverage estimation strategy for the 2021 Census of England & Wales. UKSA Census Methodological Assurance Review Panel paper EAP 112. Available at

https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP112-Over-coverage-estimation-strategy-for-the-2021-Census-of-England-and-Wales.docx

Shipsey, R. (2019) Methodology report on coverage matching for the 2021 Census. UKSA Census Methodological Assurance Review Panel paper EAP 107. Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP107-Methodology-report-on-coverage-matching-for-the-2021-Census.docx

Smallwood, S. (2020) Using demographic analysis in producing the best possible population statistics for England and Wales – Options around a national adjustment. UKSA Census Methodological Assurance Review Panel paper EAP 126. Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP126-demographic-analysis.docx

Whitworth, A., Sexton, C. and North, R. (2018) The 2021 Census Coverage Adjustment Strategy Methodological Assurance Review Panel paper EAP 106. Available at https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP106-The-2021-Census-Coverage-Adjustment-Strategy.docx

Annex A Playbook approach

We have set out, for 2021 Census, critical success factors (CSFs) that need to be met to give us confidence that we will meet end-user needs for census data and gain the benefits set out in the business case. The standard census design, including the Wave of Contact strategy, targeting of special population groups, the coverage survey and statistical processing strategies is designed to meet the CSFs.

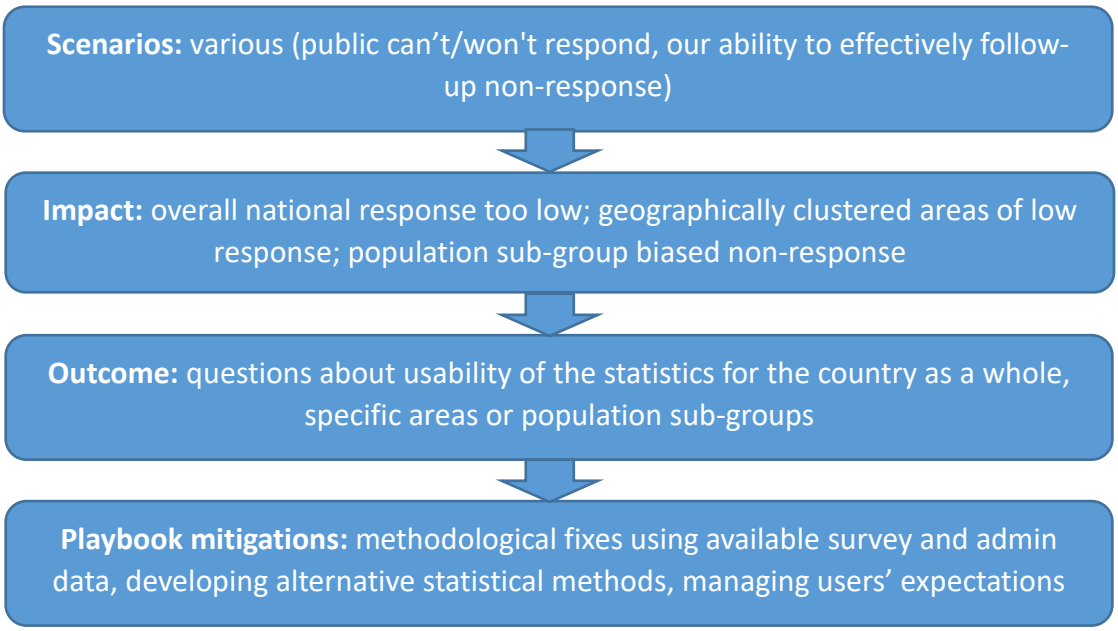
However much we are mitigating for expected eventualities during collection, previous censuses, recent international experience and the current COVID-19 situation have demonstrated that we need to be prepared for the unexpected. We are therefore setting up in advance where possible statistical contingencies that can be used in the cases of overall low response, issues in a geographic area, or in a population group. These are likely to make additional use of the information available in alternative data sources, alongside the statistical methods already in place to cope with coverage issues.

In considering our approach to these situations, we have considered:

- Possible scenarios
- Methods
- Data: including our overall administrative data strategy
- Governance
- Resources

The standard statistical design for 2021 Census includes the coverage strategy, which is there to align with our quality targets. The Playbook is a series of contingencies that can be used in scenarios where the standard design isn't sufficient to meet the targets. Figure A1 below summarises the approach.

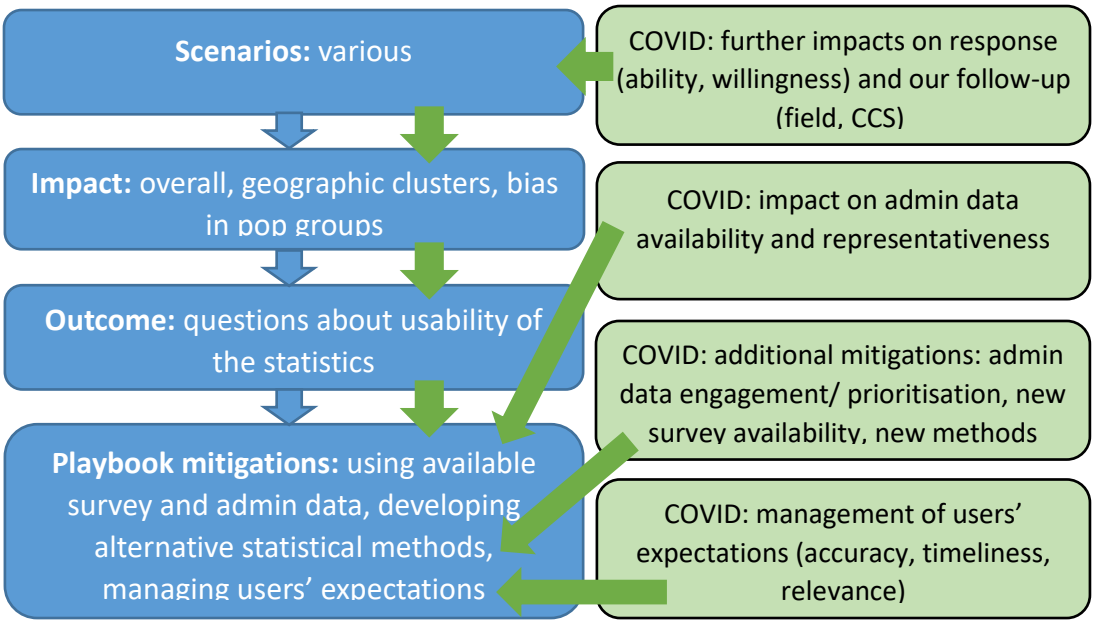
Figure A1: Playbook approach



We are also assessing what impact COVID-19 is already having, and what it is likely to have in 2021. The various knock-on effects of lockdown and the ongoing situation is likely to disrupt both standard and playbook designs, meaning that further mitigations and contingencies are required. Figure A2 below is a basic representation of the kinds of impacts we are aware of, and we will be adapting our approach as the situation develops.

Apart from the impact that the situation will have on the census collection operation, we are already facing issues of administrative data: in terms of acquisition (data suppliers affected by lockdown) and the representativeness of the data (have the population been engaging with the services these sources describe, and are the concepts they measure still relevant to our needs).

Figure A2: potential COVID-19 impact



Annex B Possible Scenarios

These are some of the scenarios we believe are (a priori) possible in 2021, and could impact the quality of the census data:

1. Overall engagement problem affecting overall response
2. Independent shock before/during collection (eg a data breach) affecting overall response
3. Flooding affecting some areas
4. Foot & mouth affecting some areas
5. Reaction/disruption to census from a particular sub-group
6. Not enough field force
7. Census Coverage Survey (CCS) response too low
8. Reaction to privacy - don't provide name/Date of Birth (DoB)
9. COVID: can't do field follow-up everywhere
10. COVID: can't do CCS everywhere
11. COVID: our collection planning assumptions are no longer relevant
12. COVID: can't get admin data/delays in admin systems catching up with changes
13. COVID: demographic/health/economic series shift
14. COVID: end data needs have changed

Table B1 summarises key ones, their likely outcomes if not mitigated, an early view of the playbook mitigation, additional COVID-19 impacts, and further mitigations or contingencies resulting from these.

Table B1: Playbook Scenarios

Scenarios	**Outcome if not mitigated**	**Greater use of admin data?**	**COVID impact / new scenario**	**Further mitigation / contingency**
Overall engagement problem (eg from data breach elsewhere)	Response rates too low for standard design to cope with	Yes	Admin sources availability and/or representativeness – can no longer rely on these sources	Prioritise alternative sources (eg utilities data), mixture of stats methods
Area specific issue (flooding, foot & mouth)	Clustered response issue in a geographic area	Yes	Admin data issues (as above)	As above
Reaction/ disruption to census from a particular sub-group	Biased response issue in specific population groups	Yes	Assumption shift - greater likelihood of this bias/variability happening Admin data issues (as above)	Possible greater reliance on COVID surveys Caveats in outputs data

Not enough field staff	Risk of increased variability by area – standard design can't fix	Possibly - reprioritise to where admin data of worst quality?	Can't do field follow-up in some places – increased risk of variability outcome. Admin data issues (as above)	As above
Reaction to privacy: public react by not providing name/DoB	Name/DoB key for matching to CCS in standard design DoB key demographic variable	Yes	May increase likelihood.	Greater reliance on alternative methods Caveats in outputs data
CCS response too low	Confidence intervals wider than CSF targets, greater uncertainty in end product	Yes	Risk that we cannot do CCS, or cannot do it everywhere, increased risk of low response rate everywhere	Need alternative, tbc
COVID changes patterns and user needs			Demographic/health/economic series shift and end-user data needs change	Mitigations tbc

Annex C Governance

As well as having alternative methods and data sources in our toolkit, one of the key lessons learned from recent international census experience in Canada and New Zealand was the governance process for spotting issues, having resource ready to investigate and develop solutions, and for an efficient governance approach around it all. We have run a series of wargames to test this, which concluded in February with assurance from our independent review panel that the following is a sound approach.

This is the governance structure:

- data quality issues are raised at Data Quality Management Forum (DQMF, meets every morning)
- issues are triaged to escalate as necessary to Stats Contingencies Extraordinary Forum (SCEF, meets if necessary, after DQMF). Triage decides if something can wait, or needs to be discussed asap
- issues that can be solved within the standard design (eg via a rule-based edit or a change to processing parameters) progress through agreed governance approach
- issues that can't be solved by tweaking standard design are discussed at more detail in SCEF. Forum recommends approach to investigating, commissions work to be done and governs the progress of commissioned task forces
- engagement with internal Census Research Assurance Group (CRAG) and external Methodological Assurance Review Panel (MARP) to get methodological assurance of the options
- once options have been explored, the recommended option is escalated to appropriate senior stakeholders' forum for decision (Processing & Outputs Delivery Forum the first step here, escalating further if necessary)

Annex D International Peer Review summary of findings

Main Points

- We conducted a war games exercise to test how our governance processes would run in live operations, under circumstances which could warrant the invoking of a statistical contingency. The review was conducted virtually for two weeks between the 8th and 18th of September 2020, followed by two shorter sessions in December 2020 and January 2021.

- The exercise was watched by a review panel as well as observers from other international organisations. The review panel was comprised of four census experts, who provided observations throughout the exercise on how the governance process worked and what could be improved upon.

- Overall feedback from the Peers from the initial exercise in September:

Overall, we observed that the governance for the management of statistical (and other) contingencies worked mostly as intended during the 'war games' simulations. We recognize that the timing and the virtual nature of the exercise may have impacted somewhat the conduct and outcome of the simulation. With seven months to go before operations, some processes and solutions are still in development, which from our professional experience is normal at this stage of preparation. The team should be commended for the organisation of the simulation and the breadth of issues developed for war gaming. They have provided an excellent opportunity to make a few adjustments to processes and governance as required before operations and further develop their playbook and risk mitigation/contingency scenario.

- Extracts from the peers' feedback after observing the December and January exercises:

… the meeting processes and flow was much improved from what was observed in round 1. The issues related to the governance process identified in the observations report in the first round were mostly addressed in rounds 2 and 3.

… also observed an improvement in the governance process between round 2 (December) and round 3 (January). Some of the uncertainties around purpose and roles expressed by some participants in triage in December were not observed in the January meetings.

Generally, the process related to each individual governance meeting worked well: the chair persons were efficient in running the meetings, the issues to discuss were well identified, and decisions and actions were generally well summarized at the end of each meeting. The introduction of an actual issue in round 3 clearly helped to focus the discussions: the conversations naturally became much more focused and decision oriented which had previously been lacking.

… . echo the marked improvements we've seen over time.

Overall – a really valuable wargame approach that I know the Stats NZ Census programme is already intending to adopt!

(Also to note that Australian Bureau of Statistics are also instigating a weekly overview of data quality, as a result of observing the exercise.)

- Next steps include
    - Continuing to do such exercises – develop a muscle-memory (key finding from the US Census Bureau's 2020 experience)
    - Agree the governance approach between collection and processing/stats contingencies – how much each can flex, understanding each other's 'playbooks'
    - Further development of methods, data and the integration of these within flexible pipelines
    - Getting buy-in from senior managers, including external assurance of proposed contingencies and live updates and challenges on the status of the data quality and timetable

Scenarios

A variety of scenarios were created by the Census Statistical Design of Processing team for the exercise. These scenarios were created to test how the individuals, who will be running live operations in 2021, would deal with a scenario, which could warrant a statistical contingency. A statistical contingency is a scenario which would force the ONS to move away from its standard design. This change from the standard design could range from changes in collection activities to standing up new teams to develop and implement larger processing changes. The final scenario tested in January was a 'live' situation: the impact of the location of higher education students on census collection and coverage survey.

The Review Panel

The review was observed by four census experts. These individuals were asked to participate in the peer review because of their extensive knowledge and experience running censuses across the globe. Furthermore, all these people are independent from the current development of the standard design which will be used in the 2021 Census. Therefore, they were able to provide us with objective feedback on the current governance processes. The reviewers who participated in the exercise were:

- Marc Hamel - Previously head of the Canadian Census and an individual with a wealth of operational experience including in contingency operations.
- Abby Morgan – An individual with a wealth of methodological expertise in the use of admin data and who was heavily involved in New Zealand's 2018 census contingency planning.
- Garnett Compton – Extensive experience with censuses across the globe and was involved in live operations during the 2011 Census. Involved in the early design of the 2021 Census, but not currently contributing to the development of the 2021 Census design.
- Professor James Brown – Heavily involved in the development and running of the ONS' census coverage methodology in 2001 and 2011. James is currently a professor at the University of Technology Sydney, and is independent from the 2021 Census design.

Key findings	**Next steps (and current progress)**
Interaction with collection • Data incidents, prioritisation, inclusion with collection incidents • Joining up playbooks	Agreement on prioritisation of issues: agreed. Worked through a range of data issue scenarios to agree routes for issues, work across teams (collection, processing, methodology) to align playbooks: being taken forward.
Stats contingencies governance during live ops • Experts want to discuss solutions but need governance to take a step back and manage, and to task expert groups to solve • Ability to take a step back and review all data issues going on, prioritise, look for patterns	Move SCEF to afternoon, to allow time to develop high level options. Done Write proposal for monthly data quality gateway/challenge sessions for live operations Proposal made for both weekly overview and monthly challenge sessions. Weekly sessions starting w/b 8 March. Monthly still in discussion.
Get some methods ready in advance • Especially covid area lockdown solutions (reality) • Subs bench earmarked and briefed (the experts we will draw on to come up with solutions) • But be ready to respond to anything live (data acquisition, pipelines not unwieldy)	Progress: Methods paper main section of this paper Subs bench skills and experience requirements drafted; next step is to get buy-in from wider ONS teams. Getting updates on data acquisition and flexibility in acquisition and internal data pipelines; need to rehearse being able to implement new data sources.
Involving methodological assurance now and live • Robust, transparent, defensible • Swift turn-around live	Engage with external methods assurance panel to agree approach for live ops – being taken forward. Discuss with Stakeholder Engagement team about other external assurance (eg expert users on high level QA panels)
Getting senior managers' buy-in • In advance – review of risks and responses • Commitment of their time in advance, and during live ops	Disseminate these findings, with concrete and directive proposal for stats contingencies governance for 2021. Get sessions into calendars of those who need to be engaged.
Keep practising, develop muscle memory • Further peer review exercises • Joining up with collection war games	Progress: further stats contingency wargames run in December, January and February. Now complete, but keeping the option to keep practising – eg holding a session on 'missing characteristics' in March.
There still tends to be too much time spent on operational [collection] contingencies. It will be an ongoing challenge for this group to stay focused on the statistical remit while the [collection] operation is still in progress.	Continue to clarify the remit of each group. Expectation that continued practice will solidify roles and responsibilities in teams.
The daily governance meeting workload is very heavy, risking key people having no time to	Being taken forward.

progress action points. Have Plan B for governance in place	

Annex E Alternative data sources including Address Centric Intelligence Datastore (ACID)

Our administrative data strategy

There is an increased emphasis on the use of administrative data in the playbook approach, but it should be noted that we already make considerable use of various sources in the end-to-end standard census design. Specific cuts of administrative data for the 2021 CEnsus

The narrative for the future (and in the context of the 2021 Census)

1. admin data will be at the heart of the statistical system going forward
2. that does not mean that a 2031 census will not be needed but that decision is down the road in 2023
3. for now, the 2021 Census is vital to us understanding the population and ever more so because of the current pandemic and our rapidly changing society.

Therefore, it is vital for everyone to take part, but it is only sensible to plan contingencies, adapt operations in a flexible way, use all available data etc. as we have done for every census.

Admin data is part of our 2021 Census strategy, following on from the National Statistician's recommendation in 2014 about using all available sources.

Publication of the statistical design is aligned with the data capability strategy, to be:

- Purposeful and confident (we shouldn't sound apologetic for using sources such as admin data as we use them responsibly and for the public good)
- Honest and open (if people think we're hiding details of data sources etc. they may not trust our outputs or share more data with us)
- Agile (we need to convey that we are taking advantage of new technologies and techniques and are continually innovating to produce the best stats)
- Collaborative (we rely on data from other organisations so it's good to acknowledge their part in our successes)
- Expert (we want to be seen as the thought leader in data sharing, data ethics and statistical outputs)

Table E1 – Main alternative sources being used in census mitigations

Data source	Comments
NHS Personal Demographic Service (PDS)	Known issues of currency of patients' locations Potentially vaccination program will improve quality
Council Tax data	Supply at risk (LA available resources), exacerbated by COVID. Indicate vacancy and wholly student households
Utilities – gas, electric	To indicate vacancy
Valuation Office Agency	To help impute type of accommodation for Dummies/AHE
Response Management data	Intelligence from collection operation
Fieldwork Management tool data	Intelligence from collection operation
2011 Census response data	Potentially impute variables that don't change
ACID (person-level address-centric admin combined intelligence data source)	See below
Alternative sources of data on HE students	Looking for alternative sources (eg direct from universities)
NHS vaccination data	Negotiating whether this source can be used for census work

ACID (Person level address centric admin + census intelligence datastore)

Summary

Intelligence from multiple sources to derive strong evidence of residency. Address-centric roster of people within addresses, with derivation of who is likely to be currently resident. Also includes census operational intelligence from Response Management and Fieldwork Management Tool (including 'dummy' forms collected by field for non-responding addresses) and address-specific administrative data sources (Council Tax, Utilities, Valuation Office Agency). Uses can include:

- Standard design:
    i. additional household structure information to feed into Coverage Estimation bias adjustment ('alternative household estimate' – also uses field dummy data and other paradata from the census collection)
    ii. additional intelligence about non-responding households to assist in placement of imputed households in Coverage Adjustment (also use field dummy data and paradata)

- Census low-response scenarios:

i. weighting classes alternative coverage estimate for non-responding households
ii. auxiliary information for use in Coverage Estimation synthetic contingency method – filling in holes at local level

Uses

We know from previous census experience both here and abroad, and from the coronavirus pandemic, that we need to be prepared for the unexpected. The Address Centric Admin Combined Intelligence Dataset (ACID) is designed to be a product that can be used in census low response contingencies. These scenarios include both national and local count issues, as well as individual questionnaires not being returned.

ACID combines various administrative and survey data sources at both individual and address levels. This combined data can then be used to provide estimates for occupied addresses, household size and household composition. In an ideal world, this product will not be necessary due to a successful census but will serve as a last resort if needed to fill any areas of weakness in the collected data. Long term, what we learn through the process of creating ACID will be used elsewhere in the office, in particular in the Population and Migration Statistics Transformation (PMST) project.

Data Sources and Summary of Approach

The ACID dataset is based on the Census address frame linked, using Unique Property Reference Number (UPRN), to the Personal Demographic Service (PDS) to populate each address with these records. The PDS contains details of all patients registered with GP practices in England and Wales and so covers much of the population the Census seeks to capture. Additionally, information about the occupants of addresses from the English School Census (ESC), Council Tax and Electoral Register datasets are included to further build up our picture of who the administrative data thinks lives in each address.

There is known to be both over- and under-coverage in the PDS from a variety of sources. The information from the other data sources about the occupants of each address is compared to each PDS record to identify which PDS records we believe are correctly allocated to that address. As our main concern is with the occupants of specific addresses, all data at this stage is linked through the UPRN.

At present, we use the responses to the Census rehearsal exercise as a source of truth to identify how accurately ACID can predict correct occupancy. These are linked, using a deterministic matching process to the individual PDS records. This allows us to flag the PDS records we believe to be correct in the address and see how this relates to the other information we have available. Using the Census rehearsal responses has meant that we have been restricted when developing our methods to just four local authorities (LAs): Carlisle, Ceredigion, Hackney and Tower Hamlets. The rehearsal exercise also had quite low response rates, only receiving back responses from 31% of addresses. As a result of these two limitations, we are currently exploring the possibility of using other surveys such as the Labour Market Survey as an additional source of truth. We would also seek to use the entire 2021 Census dataset if this were to be used in the processing and estimation if it were available in time but hope that our methods will prove robust enough with just survey data if necessary.

Modelling Summary

As discussed earlier, we are using the information provided by the various administrative data sources on who lives at a given address to predict which records from the PDS are the correct occupants of that address. To do this, we have considered the use of two different types of model; a logistic regression and a random forest. Both processes involve splitting the data into different subsets before 'training' them using one portion and then applying these learnings to test the model compared to a smaller subset of known outcomes. A series of metrics will then be used to determine which model is best and this will be constantly reviewed as more, different data is added into the model. Once a best model has been chosen, this will then be used to predict the correct occupants for any non-responding addresses.

Household Composition

The outcomes from this modelling stage will then be used to categorise addresses based on the relationships between their occupants. In this stage of the work, there will be two distinct steps. The first step will be to provide information about number of occupants and breaking these numbers down by sex and whether there are dependent children. This should be relatively straight forward and use the data directly from the PDS. More detailed analysis of the relationships between occupants will require additional data to be included at this stage and linked on to the correct PDS records we have identified through the Demographic Index (DI). The DI will enable us to link pension and benefit data which includes partner and parent/child relationships as well as incorporating Higher Education Statistics Authority (HESA) data to identify any address solely occupied by students in Higher Education. A series of deterministic rules will then be applied to this data to group addresses into the household composition categories used in standard Census analysis.

Development and Assurance Process

To this point, the development of ACID has been a Proof of Concept project and, to an extent, will continue to be so until it is decided if it is needed. Over the first half of 2021, we are going to be looking to further develop and refine our methods. This is likely to include, as stated, the inclusion of additional survey data as well as incorporating any new administrative data sources into the model if they can provide additional value.

The initial modelling phase of the project has been previously been presented to both the CDCTP Research Assurance Group (CRAG) and the Methodological Assurance Review Panel (MARP) and both broadly supportive of the approach we are taking. Comments and questions received from these groups were, for the most part, about the data we are using. The first area highlighted was about the availability of data and the fact that we do not have all data sources available for all areas as some data acquisition is done on an LA by LA basis. Their recommendation was to use as much data as possible for any LA where ACID may be used, rather than being set on one combination of data sources. Fortunately, this occurs with the four rehearsal LAs so we will be able to research and understand the impact that differing data availability has on the performance of our models. We will also be attempting to create a matrix to capture data availability by LA to be more efficient in the future.

Secondly, comments were made about data coverage and the biases that this might introduce. It is reasonable to assume that individuals who fall into hard to count groups are not likely to be as well covered by administrative data. This could be for several reasons, including having no fixed address. Due to our approach being solely reliant on the PDS and then linking other administrative data sources using address identifiers means that ACID is unlikely to be effective at capturing these groups. As we develop the methodology, we will work with colleagues across the office to try and identify any ways in which we can mitigate for the under-

representation of hard to count population. In addition, we are aware of other groups for whom we will have less data in ACID, such as children in private education as the ESC is the only 'child specific' dataset we can use. Our use of other administrative data sources and association through surname should help to mitigate the effect this has but is something to be aware of as we develop the methodology further.

Finally, the groups shared their thoughts about the timeliness of data and what any time lags and reporting dates could mean. It is well known that administrative data is not always updated promptly. As we are seeking to compare it to a fixed time source of truth, poses challenges to our methodology. Unfortunately, there is little we can do about this but we will be seeking to conduct research into the performance of various models when using different extracts of administrative data, primarily focussing on the difference between using the most recent data and the data closest to the source of truth reference date.

Moving forward, our primary source of regular feedback will be from updates to representatives from within ONS as well as from the devolved statistics authorities. We will regularly present progress updates as well as occasionally seeking guidance and answers to specific questions. In addition, we will also present our work to both CRAG and MARP as and when it is appropriate.