**Overcoverage estimation strategy for the 2021 Census of England & Wales**

**Version 1.10**

**Viktor Račinskij and Ceejay Hammond, Office for National Statistics**

**15/02/2019**

**Commented by: Owen Abbott**

1. **Introduction**

The Census aims to enumerate everyone within the population, however it can be subject to coverage error, such as undercoverage and overcoverage. Overcoverage within the census occurs when an element of interest (i.e. individual, household etc.) is either enumerated more than once, or enumerated in the wrong location, or does not exist, or does not belong to the target population. Overcoverage was not deemed a significant issue in England and Wales until the 2011 Census (Abbott and Brown, 2007). The level of overcount across England and Wales was estimated at 0.6% in 2011 (ONS, 2012a). Overcoverage was expected due to many changes in the 2011 Census from 2001, such as the increase of fieldwork (to 6 weeks), the use of online questionnaires, the option to post back (both lead to less field control and contact) and social changes, such as an increase in individuals having second residences (Abbott and Brown, 2007). Overcoverage is likely to increase further in the 2021 Census as 75% of census responses are expected to be online in 2021 (ONS, 2017) and it will be allowed to request a unique access code / a paper form to compete an individual return. Thus, it may be easier to be enumerated multiple times.

1. **2011 approach to overcoverage**

In 2001 and 2011 (3.5 weeks and 6 weeks after the Census, respectively), the Census Coverage Survey (CCS) was conducted to measure coverage across England and Wales. The CCS is carefully designed and controlled to minimise the chance of over-count. This voluntary survey, surveyed approximately 1.5% of the population (17,400 postcodes and 340,000 households in 2011), and achieved a 90.4% return rate (ONS, 2012b).

Four types of overcoverage (Abbott and Brown, 2007)

Type 1: Duplicate returns at the same location (postcode)

* + Two or more returns from the same household at the same location.

Type 2: Erroneous returns

* + These returns occur when:
		- Fictitious people are included
		- Not a usual resident of England and Wales is included
		- An individual who is born after census day or an individual that died just prior to Census night is included

Type 3: Duplicate returns from different location (postcode)

* + Two or more returns from the same individual at a different location.

Type 4: Counted in the wrong location (postcode)

* + An individual is counted in the wrong location.

These types will be reviewed for the 2021 Census.

The Dual system estimator was the under-pinning estimation strategy in the 2011 Census, as it estimates the total population with the presence of non-response (undercoverage), under certain assumptions (Brown *et al.*, 2018). This was then followed by ratio and synthetic local authority (LA) estimation.

Key assumptions of the DSE:

1. Independence between the Census and the CCS
2. Closed population (no births or deaths)
3. The probabilities of enumeration are the same for all individuals in the population
4. Perfect matching
5. No over-count

 Recall the contingency table, where cell values can be found by linking the Census and corresponding CCS:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | CCS |  |
|  |  | In | Out | Total |
| Census | In | *n11* | *n12* | *n1+* |
| Out  | *n21* | *n22* | *n2+* |
|  |  | *n+1* | *n+2* | *n++* |

$DSE=\frac{n\_{1+}\* n\_{+1}}{n\_{11}}=\frac{\left(In CCS count\right)\*\left(In Census count\right)}{Matched count}$

**2.1 An outline of the estimation of the over-count strategy in 2011 (Abbott *et al*., 2008; Large *et al*., 2011)**

*Yap* =the true count of usual residents on Census night for age-sex group (*a*) within postcode (*p*);

*Xap* = Census count of usual residents on Census night;

*Zap* = the corresponding CCS count of residents on Census night;

*Map* = the corresponding matched count between Census and CCS.

The presence of overcoverage causes the DSE to be biased as the assumption of no overcount does not hold. Therefore, the DSE needs to be adjusted for the presence of overcoverage. To adjust the DSE, the census count of usual residents on Census night is divided by the overcount propensity. The Chapman correction (Brown *et al*., 2008) is also applied to the DSE, to correct for small sample bias. So that the population total for an age-sex group *a* in a postcode ­*p* can be estimated in the following way

$$\hat{t}\_{a, p}= \frac{(Z\_{ap}+1)\*(\frac{X\_{a\_{1}p}}{\hat{γ}\_{a\_{1}}}+\frac{X\_{a\_{2}p}}{\hat{γ}\_{a\_{2}}}+1)}{M\_{ap}+1}-1$$

In 2011, it was assumed there are two different groups ($a\_{1}$ and $a\_{2}$), such as students at home addresses (we account for) and everyone else.

Where, $\hat{γ}\_{ai}=propensity=\frac{Total population count}{True population count}$ is estimated from the CCS which can detect both census duplicates and census records which are enumerated in the wrong location. However, the CCS is a sample and therefore the data could not support the estimation for a large number of population domains. Therefore, estimates were only able to be derived by broad age-sex group by hard-to-count by region. These were improved by calibrating to duplication rates which were derived from a large scale linkage study which searched for duplicates across the census data (census-to-census matching), and then sampled potential duplicates for clerical checking.

1. **An overview of the coverage estimation strategy for 2021 Census of E&W**

The 2021 Census Coverage Estimation (CCE) strategy, aims to provide high quality population estimates. The 2021 Census data and the corresponding CCS data will be linked to enable the estimation of both undercoverage and overcoverage.

Due to the increased speed in which data are expected to be received, the data for entire E&W may be available for the coverage estimation, which enables census coverage to be modelled using logistic regression or mixed effects logistic regression models (Račinskij, 2018; Alho, 1990; Alho *et al.,* 1993; Saei and Chambers, 2003; Chambers and Clark, 2012).

* 1. **Undercoverage modelling**

To obtain census population totals, a generalised linear model or generalised linear mixed model with some appropriate link function $g$ will be used. $π\_{i}$are the response probabilities. The reciprocal of these response probabilities produces the weighted Census observations.

$g\left(π\_{i}\right)=\hat{β}\_{0}+\sum\_{i=1}^{k}\hat{β}\_{i}x\_{i}$, therefore $π\_{i}=g^{-1}(\hat{β}\_{0}+\sum\_{i=1}^{k}\hat{β}\_{i}x\_{i})$, or

$g\left(π\_{i}\right)=\hat{β}\_{0}+\sum\_{i=1}^{k}\hat{β}\_{i}x\_{i}+\hat{u}\_{l}$, therefore $π\_{i}=g^{-1}(\hat{β}\_{0}+\sum\_{i=1}^{k}\hat{β}\_{i}x\_{i}$+$\hat{u}\_{l}$) where $g $is a link function.

To estimate census population totals, sum up all weighted census observations with the corresponding characteristics of interest (US Census Bureau, 2012).

This is outlined by Račinskij, (2018).

1. Generalised linear model

$\hat{T}\_{al}^{GLM}=\sum\_{r\in al}^{}\left[g^{-1}(\hat{β}\_{0}+\sum\_{i=1}^{k}\hat{β}\_{i}x\_{i})\right]^{-1}$

1. Generalised linear mixed model

$\hat{T}\_{al}^{GLMM}=\sum\_{r\in al}^{}\left[g^{-1}(\hat{β}\_{0}+\sum\_{i=1}^{k}\hat{β}\_{i}x\_{i}+\hat{u}\_{l})\right]^{-1}$

1. **Outline of overcoverage estimation strategy for 2021 Census of England and Wales**

The 2021 Census overcoverage estimation strategy will build on the 2011 approach in terms of establishing the true status of the census returns. The data needed for the purpose are not expected to be substantially different to that used in 2011 Census. No use of alternative data sources (such as admin data) is planned in the overcoverage estimation and the CCS will remain the best available source to establish the true status of the census returns. However, to account for the census overcount a different method to that used in 2011 Census is proposed. The proposed method mirrors the census undercoverage estimation and it is reasonable to expect an increase in the accuracy. This method is similar to the one used by the US Census Bureau in 2010 Census (US Census Bureau, 2008; US Census Bureau, 2012), but the way the data are provided for the estimation differs. It is likely that the estimation of the overcoverage probabilities will be carried out on a subset of the variables used in the estimation of the census undercoverage probabilities.

We are aiming to estimate the total population size of a domain of interest. This domain comprises certain individual (or household) characteristics as well as some geography attribute: say, an age-sex group *a* in an area *L* (usually a local authority). We use the vector of covariates ***x*** (which includes *a*, some other variables that do not constitute the domain, interactions of *a* with other variables) and optionally an effect of *L* in estimation.

Every element in the population has a correct location associated with it. Strictly speaking, the correct location of an element is some atomic entity such as an address. The CCS is designed to capture the correct location of an element at the census day. In what follows we assume that the CCS indeed always captures the correct location and is not subject to any form of the overcount. However, it is not always practical or necessary to define the correct location at such atomic level. We say that a search area *λ* is an area ‘around’ the atomic correct location (captured by the CCS) such that a population element with a census return within the search area is classified as counted correctly even if the atomic location is incorrect. In what follows, the search area is some small geographical area such as a postcode. But in principle, it could be entire England and Wales if all we are interested in is whether someone was counted correctly within the countries. Note that $λ⊂L.$ We write *Lλ* to denote that within the area *L* the search is carried out at the level *λ*.

Unlike the CCS, the census is prone to the overcount. We assume that only the following census related events are possible (outcome name and a symbolic representation in brackets), an element is: counted (*counted*, *I*); missed *(missed*, *Ic*); counted correctly exactly once (*correct once*, *IL1, λ*); simultaneously counted once correctly *(correct part of duplicate*, *IL2, λ*) and once incorrectly in the same or different location (*wrong part of duplicate*, *I2L*); counted once in the wrong location (*wrong*, *IcL, λ*; note the use of superscript *c* as in the missed case, this is to highlight that the wrongly counted elements contribute to the census under-count in the domains where they should have been counted). The last event, if needed, could be further split into the components: counted in the wrong location within *L* (*wrong in*, *IL, λ’*; $λ⊂λ^{'}⊂L.$, that is, an element can be found within the domain by extending the search area); counted in the wrong location outside *L* (*wrong outgoing*, *ILc, λ*; cannot find an element by extending the search area within *L*, because the element is outside *L* in *Lc*); counted wrongly in *L* (*wrong ingoing*, *ILM, µ*; an element is located in *M* within the search area *µ*, but the corresponding return is in *L*). We assume that there are no triplications, cases of being counted in the wrong location multiple times, no returns from the non-target population elements (erroneous returns). To complete the taxonomy of events: *IL1, λ* U *IL2, λ* = *IL, λ* (*correct count*).

The linkage process satisfies the 1-to-1 constraint: every CCS (census) record can be linked to one and only one census (CCS) record or unlinked. We denote the true population values by ***x*** and *L*, while those reported in the census have a superscript *r*: ***xr*** and *Lr*. We assume that in the census there are no errors reporting ***x*** (and we drop the superscript in this case)*,* but there are errors reporting *L*. An error (misreport) in *L*rλ is essentially an instance of the overcount. Under and overcount mechanisms are assumed to be independent.

When all the assumptions of the DSE are satisfied, absence of the over-count means that *L*rλ = *Lλ*andthe probability of being counted correctly in the census is π(*IL, λ* | ***x***, *Lλ*) = 1. Also, π(*I, IL, λ* | ***x***, *Lλ*) = π(*I* | ***x***, *Lλ*) = π(*m* | ***x***, *Lλ*), that is, the probability of being counted in the census and being counted correctly equals to the probability of being counted in the census, equals to the probability that a CCS record matches (*m*) a census record. Absence of the overcount also means that all census returns are correctly placed where they belong to, and the estimator for the population size of the domain *Ta, L* is

$$\sum\_{R\_{I} \in a, L^{r}}^{}\hat{π}(m | x, L\_{λ})^{-1}=\hat{T}\_{a, L},$$

where *RI* is a census return indicator. In presence of the over-count the above expression does not hold. The *wrong in* and the *wrong outgoing* events result in π(*m* | ***x***, *Lλ*) ≠ π(*I* | ***x***, *Lλ*). Whereas the duplicates, the *wrong ingoing* and the *wrong outgoing* events result in $\sum\_{R\_{I }\in a, L^{r}}^{}1\ne \sum\_{R\_{I} \in a, L}^{}1$.

To account for the overcoverage the following estimator is suggested

$$\begin{array}{c}\sum\_{R\_{I} \in a, L^{r}}^{}\frac{\hat{π}(ce | x, L\_{λ}^{r})}{\hat{π}(m | x, L\_{λ})}=\hat{T}\_{a,L}^{(oc)}\#\left(1\right)\end{array} $$

where π(*ce* | ***x***, *L*rλ) is the probability of correct enumeration in the census. Note that some alternative and equivalent (at least in principal) estimators can be considered, say,

$\begin{array}{c}\sum\_{R\_{I }\in a, L^{r}}^{}\frac{\hat{π}\left(I\_{L, λ} | x, L\_{λ}\right)}{\hat{π}\left(m | x, L\_{λ}\right)}\hat{π}\left(I\_{L, λ}∪I\_{2L}∪I\_{L, λ^{'}} \right|x, L\_{λ}^{r})\left[1+ \hat{π}\left(I\_{L, λ}, I\_{2L} \right|x, L\_{λ})-\hat{π}(I\_{L, λ}^{c}, I\_{L^{c}, λ} \left| x, L\_{λ}\right)\right]^{-1}=\hat{T}\_{a,L}^{(alt)} \#\left(2\right)\end{array}$

This estimator reflects some of the components of the over-count explicitly and also $\left[{\hat{π}\left(I\_{L, λ} | x, L\_{λ}\right)}/{\hat{π}\left(m | x, L\_{λ}\right)}\right]^{-1}= \hat{π}\left(I | x, L\_{λ}\right)$, but is cumbersome and likely to be less efficient in practice (but the same input data are needed).

Naively, the probability of correct enumeration is

$$\hat{π}(ce | x, L\_{λ}^{r})= \left(1+ \frac{\#\left[x, I\_{2L}\right]+ \#\left[x, I\_{L, λ^{'}}\right]+ \#\left[x, I\_{M, μ}^{L}\right]}{\#\left[x, I\_{L1, λ}\right]+ \#\left[x, I\_{L2, λ}\right]}\right)^{-1}$$

$$\begin{array}{c}= \left(1+ \frac{\#\left[x, wrong part of duplicate\right]+ \#\left[x, wrong in\right]+ \#\left[x, wrong ingoing\right]}{\#\left[x, correct once\right]+ \#\left[x, correct part of duplicate\right]}\right)^{-1}\#\left(3\right)\end{array}$$

where #[*y*] stands for the number of occurrences of *y*.

The overcoverage work thus aims to estimate the probability of correct census enumeration, π(*ce* | ***x***, *L*rλ). To obtain the data in (3) an additional linkage process is needed. For every CCS record this linkage process establishes whether a link can be made or not outside the search area *λ* (which includes searching inside and outside *L*). Moreover, the census to census linkage is going to carried out to detect (but not resolve) duplicates. It is yet to be decided how to use the data from the latter exercise.

The ongoing overcoverage work focuses on: empirical checks of (1) and specification of the data needed in estimation / outlining processes that are required to derive these data; clarification of the underlying assumptions; establishing the basic statistical properties of (1) using simple simulation studies.

Future work will attempt to model the probability of correct census enumeration using 2011 Census data; clarify the level of model complexity supported by the available data; understand the effect of sampling design on the undercoverage estimation; understand the effect of the overcoverage on the undercoverage estimation; decide how the data from the census to census linkage exercise can be used; conduct the full-scale simulation studies with under- and overcoverage.

**References**

Abbott, O., and Brown, J., (2007) ‘Overcoverage in the 2011 UK Census’. Paper presented to 13th Meeting of the National Statistics Methodology Advisory Committee.

Abbott, O., Brown, J. and Skinner, C., (April 2008). Response to discussion at NSMAC on estimation of over-count.

Alho, J. (1990). Logistic Regression in Capture-Recapture Models. *Biometrics*, 46, 623-635.

Alho, J., Mulry, M., Wurdeman, K. and Kim, J. (1993). Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. *Journal of the American Statistical Association*, 88, 1130- 1136.

Brown, J. J., Sexton, C., Abbott, O., & Smith, P. A. (2018). The framework for estimating coverage in the 2011 Census of England and Wales: combining dual-system estimation with ratio estimation. To appear in Statistical Journal of the International Association for Official Statistics.

Chambers, R. L. and Clark R. G. (2012) *An Introduction to Model-Based Survey Sampling with Applications*, Oxford University Press, New-York, USA.

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

Office for National Statistics (2012a). Overcount Estimation and Adjustment.

Office for National Statistics (2012b) 2011 Census Coverage Survey Summary (2011 Census: Methods and Quality Report). Office for National Statistics. Available from <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/census-coverage-survey-summary.pdf> [Accessed 9 Oct 2018].

Office for National Statistics (2017) 2017 Census Test Report. Available from <https://www.ons.gov.uk/census/censustransformationprogramme/2017censustest/2017censustestreport> [Accessed 15 February 2019].

Račinskij,V., 2018., Coverage Estimation Strategy for the 2021 Census of England and Wales.

Saei, A. and Chambers, R. (2003) Small area estimation under linear and generalized linear mixed models with time and area effects. *S3RI Methodology Working Papers M03/15,* Southampton Statistical Science Research Institute. Available from <https://eprints.soton.ac.uk/8165/1/8165-01.pdf> [Accessed 15 Jan 2019].

US Census Bureau (2008). 2010 Census Coverage Measurement Estimation Methodology. US Census Bureau, Washington, D.C. Available from

<https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf> [Accessed 21 Sept 2018].

US Census Bureau (2012). 2010 Census Coverage Measurement Estimation Report: Aspects

of Modeling. US Census Bureau, Washington, D.C. Available from

<https://www.census.gov/coverage_measurement/pdfs/g10.pdf> [Accessed 14 Jan 2019].