Small Area Estimation Strategy for 2011 UK Census

Baffour, Bernard
Southampton Statistical Sciences Research Institute
Univeristy of Southampton, UK

Silva, Denise
Brazilian Institute of Geography and Statistics - IBGE
Brazil
E-mail: denise.silva@ibge.gov.br

Taylor, Alan and Sexton, Christine
Office for National Statistics - UK
E-mail: alan.taylor@ons.gov.uk

Veiga, Alinne
Brazil

Introduction

A Census has to produce accurate and reliable estimates of the population, not just at the national level but also at lower geographical detail. However, it is also widely known that despite all the efforts of the census, there will be some people missed. The final population estimates have to be adjusted for this undercount. In the UK Census at a national level, this is achieved through a combination of a post-enumeration survey, the Census Coverage Survey (CCS), and two statistical methodological approaches known as Dual System Estimation (DSE) and Ratio Estimation. The CCS is an intensive re-enumeration of a selected sample of the UK population that takes place shortly after the Census has been completed. After matching the Census and CCS databases together, it becomes possible to find an estimate of the coverage of the Census. The final population estimates are then revised taking this coverage adjustment into account.

In the design of the CCS, the UK is divided according to a broad regional classification. These areas are referred to as Estimation Areas (EAs) that are mutually exclusive groups of Local Authorities (LADs). To obtain LAD estimates of the population, Small Area Estimation techniques that rely on direct or indirect information from the CCS has to be used. This paper reports the research undertaken for developing the small area methodology for the 2011 UK Census. A very simplified description of the UK Census Coverage Assessment and Adjustment process identifies the following key elements in the process: the Census itself; the CCS; the matching of Census and CCS databases to estimate undercount at national and sub-national levels; **the process to obtain model based population estimates for Local Authorities**; the production of a database with individual and household level records consistent with LAD population level estimates.

For the 2011 UK Census, the CCS is a nationally representative sample of 375,000 households (grouped into postcodes which are small geographical units made up of 30 households on average). On completion of the CCS, two main phases of estimation will be carried out; firstly the EA population totals are calculated through DSE. In the sampled postcodes, estimates of undercount are obtained by matching the Census and CCS records. To compute undercount estimates of the whole EA, a simple ratio estimator is used based on the dual system estimates and the Census counts. In the second phase, the LAD estimates are derived using the relationship between the initial Census count and the CCS count adjusted for the level of undercount. The EA population estimates can be directly derived using the CCS information. On the other hand, the LAD population estimates need to be found using indirect information through small area estimation. Direct estimation of LAD population using sample-specific information from each LAD is possible, but it is less precise, and thereby has large standard errors.

In 2011, the UK Census generally follows the framework of the 2001 One Number Census (Abbott, 2009). A more detailed explanation for the ONC methodology can be found in Brown et al. (1999). The purpose of this work was to consider a wide range of small area models to support the decision regarding the most suitable strategy to be implemented in 2011. A series of simulation studies were carried out to assess the performance of these models.

Small Area Estimation Models

The main objective of the small area estimation strategy is to produce reliable population estimates, and corresponding precision measures, by age-sex groups and hard-to-count[1](HtC) strata in each LAD. The small area estimation procedure apportions the sub-national EA estimates to the LADs by assuming relationships between the undercount pattern at the LAD level and the broader areas (i.e. the EAs). The underlying idea of small area estimation is to exploit similarities in order to borrow strength over areas. Regression models that relate the CCS and Census count are used for describing the relationships so as to produce the LAD model-based estimates. However, although more precise, the resulting model-based estimates are biased. Various regression models were then considered and the objective of the small area estimation procedure was to find the estimator that balances the trade-off between variance and bias, yielding estimates with good precision and as little bias as possible.

Simulation studies were used to investigate the different small area estimation methods. One of the techniques tested was the local fixed effects model implemented for the 2001 Census. In this approach, models containing fixed area effects are fitted separately in each EA. A broader area fixed effects model was also considered in which small area estimation models are fitted within groups of EAs. The Government Office Region[2] level (GOR) was chosen to define the broader areas for estimating the regional fixed effects models. In addition, regional mixed models, in which the LAD effects are random, were also evaluated.

Small Area Estimation for Local Authorities Districts

Seven different small area models were investigated (Baffour at al., 2010). These models are the direct estimator, the synthetic estimator and a number of different regression models with LAD and age-sex effects specified as either random or fixed effects. This paper only presents simulation results for the ***Direct***, ***Synthetic*** and ***Local Fixed Model*** estimators since the results indicate that the last two constitute good options to produce accurate and reliable LAD population estimates.

The age-sex categories used were defined according to 35 groups given by males and females under 1 year old, males from 1 to 4 year old, females aged 1 to 4 years old, then 5 year age groups. Most of the small area estimation techniques were tested using a set of collapsed age-sex groups for computing the estimates. The 35 groups were collapsed into 16 groups (each age-sex group c is defined such that $a \in c$).

Let $Y_{kad\lg r}$ be the DSE count for postcode *k*, age-sex group *a*, HtC stratum *d*, Local Authority *l* in a given Estimation Area *g* and Government Office Region *r*. Also, let $X_{kad\lg r}$ be the corresponding unadjusted Census count. The general objective is to produce model-based estimators for the population total by LAD, HtC stratum and age-sex group, T_{adl}. The model-based estimates $\hat{T}_{adl}$ are scaled to the EA age-sex population total, $\hat{T}^{g}_{ad}$. This calibration ensures that estimates produced by the small area (i.e. the LAD) modelling are consistent with the larger area (i.e. the Estimation Area) population estimates.

The ***direct estimator*** assumes that the sampled postcodes within the LAD have the same undercount as that which is observed in the whole LAD at a given HtC stratum, i.e. non-sampled postcodes behave similarly to those sampled in the CCS. The population of the LAD is found based on the ratio of the adjusted DSE count and the census count, and the ratio is smoothed by using the collapsed age categories *c*. This explicit use of information sets the *direct estimator* apart from the indirect models that use implicit information to describe the relationships. The *direct estimator* only uses data from postcodes in the specific

[1] The Hard-to-Count Index identified each postcode as either ‘easy’, ‘medium’ or ‘hard’ based on demographic, social and economic characteristics. These HtC strata were in the design of the CCS.

[2] For the definition of Government Office Regions consult http://www.statistics.gov.uk/geography/gor.asp.

LAD borrowing strength over sampled postcodes in each LAD. The population estimate for age-sex group *a*, in HtC stratum *d* and LAD *l* in a given Estimation Area is calculated as

$$\hat{T}_{adl} = \left[\sum_{s} Y_{kcdl} \Big/ \sum_{s} X_{kcdl} \right] \sum_{g} X_{kadl} = \hat{\theta}_{cdl} \sum_{g} X_{kadl}$$

where *s* represents summing across the postcodes in the sampled areas, and *g* represents summing across the postcodes in both the sampled and non-sampled areas. Distinct LADs within the EA will have different adjustment factors.

The ***synthetic estimator*** uses data from all the LADs within a specified EA. The underlying assumption is that the LADs have the same undercount pattern that is observed in the whole EA and that the non-sampled postcodes exhibit the same behaviour as the CCS-sampled postcodes. This estimator uses the level of undercount in each age-sex category by HtC stratum in the EA to adjust the LAD census populations. Unlike the *direct estimator*, a model is now used to obtain the adjustment factors, $\hat{\theta}_{adl}$. This is accomplished by a simple linear regression model through the origin with the DSE-adjusted counts as the response variable and the unadjusted Census counts the explanatory variable. In order to incorporate the age-sex differences each age-sex group is assigned a different slope coefficient. The model is fitted separately for each HtC stratum within each EA using age-sex by postcode level data. Thus, for the *synthetic estimator*, the model is given by: $Y_{kadl} = \theta_{ad} \; X_{kadl} + \varepsilon_{kadl} \sqrt{X_{kadl}} \quad with \quad \varepsilon_{kadl} \sim N\left(0, \sigma_d^2\right)$

$$Var\left(Y_{kadl} \mid X_{kadl}\right) = \sigma_d^2 X_{kadl} \quad and \quad Cov\left(Y_{kadl}, Y_{ja'd'l'} \mid X_{kadl}, X_{ja'd'l'}\right) = 0 \quad for\ all\ k \neq j \ \ and \quad a, a', d, d', l, l'$$

The *synthetic estimator* for the population total by LAD, HtC stratum and age-sex group is defined as $\hat{T}_{adl} = \sum_{k} \hat{\theta}_{ad} \; X_{kadl}$ where s_{dl} is the set of sampled units in HtC stratum *d* and LAD *l*.

The ***local fixed effects model*** is another indirect estimator similar to the *synthetic* one as a simple linear regression model is fitted that relates the DSE-adjusted counts with the unadjusted Census counts. The main difference is that the regression coefficients are allowed to vary according to the LADs. Again the model is fitted to each HtC stratum within each EA using age-sex by postcode level data. In each EA, the model specification is given by: $Y_{kadl} = (\theta_{cd} + \gamma_{dl}) \; X_{kadl} + \varepsilon_{kadl} \sqrt{X_{kadl}} \quad with \quad \varepsilon_{kadl} \sim N\left(0, \sigma_d^2\right)$

$$Var\left(Y_{kadl} \mid X_{kadl}\right) = \sigma_d^2 X_{kadl} \quad and \quad Cov\left(Y_{kadl}, Y_{ja'd'l'} \mid X_{kadl}, X_{ja'd'l'}\right) = 0 \quad for\ all\ k \neq j \ \ and \ \ a, a', d, d', l, l'$$

with the collapsed category levels satisfying $a \in c$ and the LAD effects in each EA are assumed to sum to zero, $\sum_{l \in g} \gamma_{dl} = 0.$ The model-based estimator for the population total by LAD, HtC stratum and age-sex group is defined as $\hat{T}_{adl} = \sum_{k} \left(\hat{\theta}_{cd} + \hat{\gamma}_{dl}\right) X_{kadl}$.

This model includes an overall age-sex effect and a LAD specific effect. In addition random error terms may differ by age-sex, HtC, LAD and EA. The model specification reflects the CCS survey design and accounts for the differences in slope terms between LAD and HtC strata.

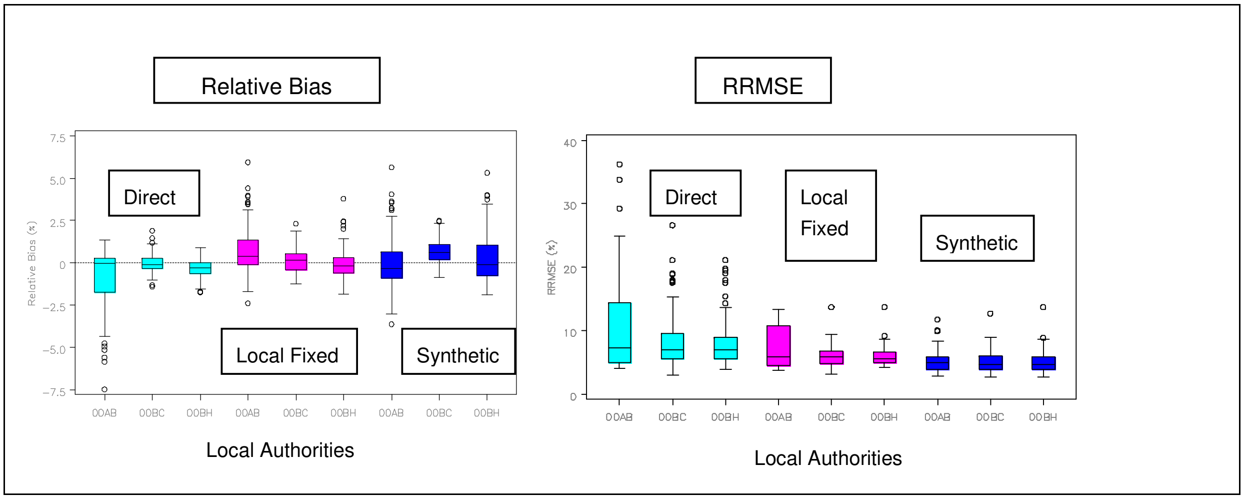
Evaluation of the Small Area Methods and Results of the Simulations

Different Census and CCS responses were simulated. This was done based on data from the Census and CCS response rates in the 2001 Census. Based on data from the previous census, 400 'Censuses' and 400 'CCSs' were simulated. For each Census-CCS pair, the coverage assessment was carried out to estimate the EA population totals through DSE. On producing these EA totals, the LAD population totals were found. For each simulation, different small area models were compared by producing LAD estimates by age-sex group and HtC stratum for each of the 400 simulations. The *relative bias* and *relative root mean squared error* are suitable measures of performance that were used to investigate the bias and variance. The aim was to find the small area estimation strategy that is robust under the different simulated scenarios.

Simulated Census and CCS data were obtained for some EAs which were selected because they had different levels of coverage in the 2001 Census. The investigation sought to determine how each of the different small area models fared under a range of coverage scenarios. Seven EAs were chosen because they represented a cross-section of areas exhibiting diverse census coverage characteristics and coverage rates varying from 76.6% to 97.7%. For each of the EAs, the RRMSEs and RBs were calculated for competing small area estimation techniques. Boxplots are used to present the distribution of the RRMSEs and RBs in order to provide information on the behaviour of the different small area estimators under the simulations. The small area technique that performs well should produce an RRMSE distribution with lower medians and a smaller spread. In the case of bias, a good technique should produce an RB distribution that is centred around zero with small spread.

As an example, the results for one EA (*Outer London*) are graphically displayed. Figure 1 shows the distribution of the RBs and RRMSEs for the 105 (i.e. 35 x 3) age-sex by HtC model-based population estimates for each LAD providing evidence that the *synthetic estimator* performs best in comparison to the *local fixed model* and the *direct estimator* when considering the RRMSEs. In general, the RRMSEs of the *synthetic estimator* are lower. Furthermore, the distributions have smaller spread within LADs. However, when using the RBs, the *local fixed model* produces better-behaved distributions, which are mostly centred around zero and are therefore fairly unbiased. The reasoning behind the *local fixed model* is to capture any difference in coverage due to LAD effects. Although no improvement in regards to the RRMSE was found, the model containing LAD effects might protect the estimation procedure against failure if LAD differentials were observed in the 2011 Census. This motivates the use of the *local fixed model*, especially in areas of poor coverage.

Figure 1 – Outer London Estimation Area - RB and RRMSE Distribution of the Estimators



Another performance measurement is the proportion of times each of the three small area estimators produces the closest estimate to the true population total. In reality, the LAD population total $T_{ad\lg}$ is not known in advance, but a model-based estimate $\hat{T}_{ad\lg}$ can be found. The measurement is given by $P_m = \frac{1}{400}\sum_{j=1}^{400} I\left(min\left[\left|\hat{T}_{ad\lg\, jm} - T_{ad\lg}\right|\right]\right)$ where $T_{ad\lg}$ is the true population count for the age-sex group *a*, HtC stratum *d* and LAD *l* and $\hat{T}_{ad\lg\, jm}$ is the model-based population estimate model *m* obtained from the j^{th} simulation, with $j = 1, \ldots, 400$, and $\mathrm{I}(\;)$ is an indicator representing whether or not the small area model *m* produces the closest estimate (in absolute terms) to the true population value. Table 1 provides an indication to which method gives the best estimate the majority of the time. The *synthetic estimator* produces the most 'hits' (those in which the estimates are closest to the true population total the majority of the time).

The simulation results indicated that both the *synthetic estimator* and *local fixed model* are reasonable options to produce LAD population estimates. The first performs better in terms of RRMSE whereas the latter produces estimates with smaller biases. The *synthetic estimator* seems more stable as it shows less variability in performance across LADs. The use of a *local fixed model* can represent a safeguard for the LAD undercoverage differentials that may be actually detected in the 2011 Census. However, the *local fixed model* may add unnecessary noise into the estimates if there are no LAD effects.

One compromising solution was to implement a small estimation procedure that could accommodate both options. That is, for each EA the procedure fits a *local fixed model*, then tests the significance of the area effects and automatically fits the *synthetic estimator* if the Local Authority effects are not significant.

Table 1 - Proportion of times the Small Area Estimation Model yields the closest estimate of the simulated population for Inner and Outer London Estimation Areas

Estimation Area	**Local Authority**	**Small Area Estimation Models**		
		Direct	**Synthetic**	**Local Fixed**
LB	00AM	0.17	**0.51**	0.33
	00AU	0.31	**0.35**	0.34
	00BG	0.30	**0.46**	0.24
LJ	00AB	0.23	**0.51**	0.26
	00BC	0.20	**0.60**	0.21
	00BH	0.19	**0.59**	0.22

Diagnostic Measures of Fit for the Synthetic Estimator and the Local Fixed Model

Although the simulations have been looking at 400 different Census and CCS combination, in reality there will only be one Census and CCS. Therefore, by looking at the diagnostic measures of fit – here the Adjusted R^2 values and the proportion of residual outliers – it is possible to provide some insight as to possible model failure and how well the different models perform. An investigation of the outliers is important in ensuring that the modelling procedure is robust. For both diagnostic measures (R^2 and outliers), the focus was to compare the *synthetic estimator* and the *local fixed model*.

The first point to make is that, in the main, the mean and median R^2 values were above 0.90 in all the EAs. The interpretation is that, for both the *synthetic* and *local fixed* indirect estimates, the models explain approximately 90% of the total variation. Secondly, there is not much difference in the R^2 values of the *synthetic* and *local fixed models*; though the *local fixed model* performs better than the *synthetic estimator* in terms of R^2. Thirdly, and possibly in support of the design of the CCS, the R^2 reduces with enumeration difficulty, i.e. an increase in the HtC level leads to a reduction in the R^2 . Regarding the outliers, the *local fixed model* appears to have larger proportion of outlying observations than the *synthetic estimator*. In most large samples, as a rule of thumb, roughly 5% of the residuals are expected to be outliers. For the *local fixed model*, the mean number of outliers obtained was greater than 5%. This implies that there are a greater number of extreme differences between DSE results and model-based estimates under the *local fixed model* than would be expected. The *synthetic estimator*, on the other hand, does not show this problem.

The number of times the LAD effects are significantly different from zero was also investigated. In most cases, more than 50% of the simulations have significant LAD effects. In a direct comparison, it would appear that the *synthetic estimator* performs better than the *local fixed* one. This does seem counter-intuitive, but bearing in mind that the *synthetic* model explains a considerable amount of the total variation, there is not much difference between the estimates derived under the *synthetic* and *local fixed* model.

Conclusions

Indirect models can be used to improve the precision of the LAD estimates. The *synthetic estimator* is often the most appropriate indirect model. The reason for this could be attributed to the design of the CCS in that the broad stratification of the UK into EAs appropriately groups similar LADs together. However, when there is localised failure of the Census (and/or CCS) - for example, a specific LAD behaves differently to the EA within which it is found - then the *synthetic estimator* can be less precise than the *local fixed model*. When it is required to choose between the *synthetic* estimator and the *local fixed* model, the model diagnostics show that the *synthetic* estimator performs better than the *local fixed*. On the other hand, the ability of the *local fixed model* to cope with LAD differentials can be its strength when there are unexpected outcomes. Thus the recommendation made from these simulation results is that the most appropriate small area strategy is to have a modelling procedure that accommodates both *synthetic* estimation and *local fixed* effects regression. The *synthetic estimator* can be thought of as the default technique of choice, and can cope with some Local Authority differentials. However, in the case that there are unanticipated problems in the Census and the CCS leading to greater differences in the observed Local Authority coverage levels, then the *local fixed model* may be better placed to produce more robust population estimates.

REFERENCES

Abbott, O. (2009) 2011 UK Census Coverage Assessment and Adjustment Methodology. Population Trends, 137 (Autumn edition), 25-32. Available at http://www.statistics.gov.uk/downloads/theme_population/PopTrends137web.pdf

Abbott, O., Brown, J., Chambers, R. and Cruddas, M., (2000a). One Number Census Local Authority Estimation. Paper submitted to the One Number Census Steering Committee numbered as ONS(ONC(SC)00/03B)). Available at http://www.statistics.gov.uk/census2001/pdfs/sc0003b.pdf

Abbott, O., Brown, J., Chambers, R. and Cruddas, M., (2000b). One Number Census Estimation Update. Paper submitted to the One Number Census Steering Committee numbered as ONS(ONC(SC)00/16). Available at http://www.statistics.gov.uk/census2001/pdfs/sc0003b.pdf

Baffour, B. , Silva, D. , Sexton, C., Veiga, A. (2010). Small Area Estimation Strategy for the 2011 Census. Unpublished ONS report. 110p.

Brown, J.J., Diamond, I.D., Chambers, R.L., Bucker, L J., Teague, A. D. (1999). A methodological strategy for a one-number census in UK. Journal of the Royal Statistical Society: Series A, 162, Part 2, pp.247-267.

Cruddas, M. (2001) One Number Census Methodology. Paper submitted to the One Number Census Steering Committee numbered as ONS(ONC(SC))01/01. Available at http://ww.statistics.gov.uk/census2001/pdfs/sc0101.pdf

ONS (1999). The role of Dual System Estimation in the 2001 Census Coverage Surveys of the UK. One Number Census Steering Committee paper 99/07. Available at www.statistics.gov.uk/census2001/pdfs/sc9907.pdf.

ONS (1999). A Guide to the One Number Census (ONC) http://www.statistics.gov.uk/census2001/pdfs/oncguide.pdf

Steele, F., Brown, J., Chambers, R., (2002). A controlled donor imputation system for a one-number census. Journal of the Royal Statistical Society: Series A, 165, Part 3, pp.495-522.

ABSTRACT

A Census has to produce accurate and reliable estimates of the population, not just at the national level but also at lower geographical detail. However, it is also widely known that despite all the efforts of the census, there will be some people missed. Therefore, the final population estimates have be adjusted for this undercount. In the UK Census at a national level, this is achieved through a combination of a post-enumeration survey, referred to as the Census Coverage Survey (CCS), and two statistical methodological approaches: the Dual System Estimation and Ratio Estimation. In addition, to obtain lower level estimates of the population, Small Area Estimation techniques are employed. The use of model based small area estimation methods in the UK Census was introduced in 2001 within the One Number Census (ONC) methodology. The same approach will be adopted for the 2011 UK Census however new features were tested. The methodology aims at obtaining local level population estimates by age-sex groups adjusted for undercount. The UK Census process includes: the Census itself; the Census Coverage Survey (CCS); the matching of Census and CCS; the use of Dual System and ratio estimation to estimate undercount; the process to obtain local area model based population estimates and the production of a census database with individual and household level records consistent with the estimates.

Small area estimation techniques based on regression models are being employed to produce local level estimates using data from the CCS and the Census. The idea of small area estimation is to exploit similarities in order to borrow strength over areas. Regression models that relate the CCS and Census count are then used as a tool for describing the relationships so as to produce the local level model based estimates. However, although more precise, the resulting model-based estimates are biased. In this study, various regression models were considered and the objective of the small area estimation procedure was to find the estimator that balances the trade-off between variance and bias, yielding estimates with good precision and as little bias as possible. This paper reports the research undertaken for developing the small area methodology for the 2011 Census, the framework for evaluation and corresponding results.