



ONS(ONC(SC))00/16

## ONE NUMBER CENSUS STEERING COMMITTEE

### One Number Census Estimation Update

1. This paper reports the additional research towards developing a methodology for estimating the populations of Design Groups and Local Authority Districts as part of the One Number Census. It is essentially a continuation of Steering Committee papers ONC(SC)00/03A and ONC(SC)00/03B. The following were researched:
  - a) Dependence between the Census and CCS;
  - b) Alternatives for prediction outside of the sample range;
  - c) Variance Estimation methodology;
  - d) The agegroups used in the LAD estimation models;
  - e) Composite LAD estimators; and;
  - f) Application of LAD estimation methodology to an area of Inner London.
  
2. **Members of the Steering Committee are asked to:**
  - a) **Note the results presented in the paper;**
  
  - b) **Agree the recommendations that:**
    - **As the performance of the two strategies for dealing with prediction outside the sample range do not differ, because of its intuitive appeal the revised approach is preferred;**
    - **a Jackknife variance estimator is used for the estimation of variances associated with the Design Group population estimates; and;**
    - **the use of collapsed agegroups should be considered further for the Local Authority District estimation.**

Owen Abbott  
Census Division, Room 4200W  
Office for National Statistics  
Segensworth Road  
Titchfield  
Fareham  
Hampshire  
PO15 5RR

## EXECUTIVE SUMMARY

### A Design Group Estimation

#### Dependence between Census and CCS

Further simulation work was undertaken to examine the impact of dependence between the Census and CCS on the ONC population estimates. These simulations demonstrate that dependence is important. However, for even quite extreme levels of dependence, the impact is small provided both the Census and CCS have high response rates.

As the response rates fall, the estimation strategy becomes more susceptible to the effects of dependence. It is worth noting that in the situation of odds ratios greater than one (intuitively the most likely scenario), the estimation strategy will, in terms of bias, always do better than the unadjusted census.

#### Prediction outside of the sample range

Two alternatives to prediction outside of the sample range were implemented within a set of simulations to compare their performance. As the simulations not conclusively favour either **approach it is recommended that the revised strategy (robust ratio 2) for dealing with prediction outside the sample range is the preferred approach due to its intuitive appeal.**

#### Variance Estimation

Two approaches to the problem of variance estimation were compared. The comparison indicated that the jackknife variance estimator tracks the true variance more closely. **It is therefore recommended that a jackknife variance estimator is used for the estimation of variances** associated with the ONC Design Group population estimates.

### B Local Authority District Estimation

#### The agegroup categories used in the LAD estimation models

The approaches evaluated in previous research used a collapsed set of agegroups. These were compared with the results from the same approaches using uncollapsed categories. It clearly shows that greater precision was obtained by the collapsing of the agegroups.

**Therefore the results of the previous research still hold. Furthermore, it is recommended that the collapsing of agegroups should be considered further for the ONC Local Authority District estimation.**

## **Composite Estimators**

Research was carried out to examine whether any gains from using a composite estimation approach could be made. Two alternative composite estimators were implemented within a set of simulations in order to compare their performance with the simple synthetic model and LAD adjusted synthetic model. The results clearly show that the composite estimators are not significantly more efficient than the LAD specific fixed estimator. Even though gains are made, there are added complications of variance estimation when using such estimators.

**It is therefore concluded that composite estimators do not provide enough gains in efficiency to warrant their use.**

## **Inner London**

A simulation study was undertaken using 1991 Census data for a Design Group within Inner London to provide more information on the robustness of the recommended strategy. The results show **the recommendation to use the fixed LAD effect model still holds.**

# ONE NUMBER CENSUS ESTIMATION UPDATE

Owen Abbott, James Brown, Ray Chambers and Marie Cruddas.

## 1. INTRODUCTION

1.1 This paper is essentially a continuation of Steering Committee papers ONC(SC)00/03A and ONC(SC)00/03B. The research within this paper is a result of the outcomes from the Steering Committee's discussion of the aforementioned papers, together with the further work that was noted within the papers themselves. It will therefore not include a detailed introduction or background to the overall aim of the One Number Census (ONC) estimation project. The paper assumes that the reader is familiar with the previous ONC estimation papers.

1.2 The paper is split into two main sections – Design Group Estimation and Local Authority District (LAD) Estimation. Within each of these the further research is reported, and recommendations made. The composition of the paper is as follows:

### **Design Group Estimation**

- Sensitivity of the adjusted ratio estimator to dependence between the Census and CCS
- Sensitivity of the adjusted ratio estimator to the choice of adjustment for prediction of large census counts
- The method of variance estimation

### **Local Authority District Estimation**

- The agegroups used in the LAD estimation models
- Composite LAD estimators
- Inner London Simulations

1.3 For each subsection, the further research is described and the results presented. Each of the two main sections provides the evidence necessary to recommend the final approach. The final section of the paper outlines the overall recommended strategy for the production of the One Number Census population estimates, and any further work required.

## 2. DESIGN GROUP ESTIMATION

### 2.1 Sensitivity of the adjusted ratio estimator to dependence between the Census and CCS

#### **Introduction**

2.1.1 Simulations in Steering Committee paper ONS(ONC(SC))00/03A assumed independence between the Census and CCS counting processes. This is a working assumption that is unlikely to be true in the field. Therefore, it is

important to demonstrate the impact of dependence on any proposed estimation strategy. Some initial work on this is presented in Brown *et al* (1999) but this only looked at using simple regression estimators. It showed that the regression estimator still performs well in the presence of correlated non-response. As the CCS response rate decreases, the direction and extent of the dependence become important, especially for estimating large census undercounts e.g. young males. The estimation strategy has developed considerably since then, although one would expect the impact of dependence to be similar.

- 2.1.2 This research investigates the impact of dependence on the estimation strategy proposed in paper ONS(ONC(SC))00/03A through the use of a simulation study. The simulations will use extreme odds ratios to examine the impact of dependence for different levels of coverage. The minimisation of the operational dependence is described in ONS(ONC(SC))00/13.

### **Simulation Methodology**

- 2.1.3 In general, the simulations assume that at the population level the census coverage is approximately 95% and the CCS coverage is approximately 88%. If the two counts are independent (the assumption made in previous simulations) then the theoretical bias of the DSE due to dependence is zero, you would 'expect' the estimator to give the correct answer. However, under the stated coverage assumptions, the bias in the DSE due to dependence could range from -5% (the CCS finds no new people – odds ratio  $\Psi \infty$ ) through 0% (the CCS is independent – odds ratio = 1) to +0.7% (the CCS finds all the missed people – odds ratio = 0). In other words, the expected value of the estimator will be between 95 per cent and 100.7 per cent of the true value, according to the exact nature of any dependence. For high response rates the impact of dependence is limited, particularly in terms of the potential positive bias. However, if census coverage remains unchanged, but the CCS coverage drops to 75% the potential positive bias increases to +1.8%. This demonstrates, in theory, the importance of high CCS coverage. The potential negative bias is limited by increasing census coverage.
- 2.1.4 Previous work introduced dependence into the simulations through specifying a specific odds ratio between the census and CCS. The choice of this odds ratio is rather arbitrary, previous work used 10 and 0.1. However, the following results compare odds ratios of 8 (people counted in the census are more likely to be counted by the CCS) and 1/8 (people missed in the census are more likely to be counted by the CCS) with the independence scenario. These are chosen as extreme values (representing  $\log_e$  odds of 2 and -2). An odds ratio of 8 is also close to the situation (with census coverage of 95% and CCS coverage of 88%) where the theoretical bias is -2%, in other words the population estimate gets within 2% of the true population.
- 2.1.5 **Table 1** below illustrates the case where the odds ratio is 8.

**Table 1: True Population with Census coverage=95%, CCS coverage=88% and odds ratio= 8**

	<b>In CCS</b>	<b>Missed CCS</b>	<b>TOTAL</b>
<b>In Census</b>	0.854	0.096	0.95
<b>Missed Census</b>	0.026	0.024	0.05
<b>TOTAL</b>	0.88	0.12	1

Odds ratio is not exactly 8 due to rounding errors

2.1.6 The estimate of missed by both through the application of DSE technology will be 0.003 ( $0.026 \times 0.096 / 0.854$ ), which is 1/8 of the true number in Table 1. Equivalently, the DSE will underestimate those missed by both the census and CCS by 7/8. However, that does not mean it will underestimate census underenumeration, in this case five per cent, by 7/8 as the CCS will still find some of the missed people. The DSE will actually estimate the total population as 0.979. Therefore, for the specific census and CCS coverage's stated above, an odds ratio 8 is not only the point where the DSE underestimates those missed by both by 7/8, but also corresponds to an overall bias in the DSE of -2 per cent.

## Results

2.1.7 Results from the simulations are presented in what follows. When considering dependence it is its impact on the bias of the DSE that is particularly important. As this is a simulation study the truth is always known. Therefore, over the simulation the bias, relative to the truth, of any estimator can be calculated as

$$\text{relative bias} = \frac{100}{\text{truth}} \sum_{i=1}^{1000} (\text{estimator}_i - \text{truth})$$

where  $\text{estimator}_i$  is the value for the estimator calculated from the  $i^{\text{th}}$  iteration of the simulation. The root mean square error (RMSE) is also important as a measure of the estimator's total error (bias and variance) under different dependence scenarios. Relative RMSE is calculated as

$$\text{relative RMSE} = \frac{100}{\text{truth}} \sqrt{\sum_{i=1}^{1000} (\text{estimator}_i - \text{truth})^2}$$

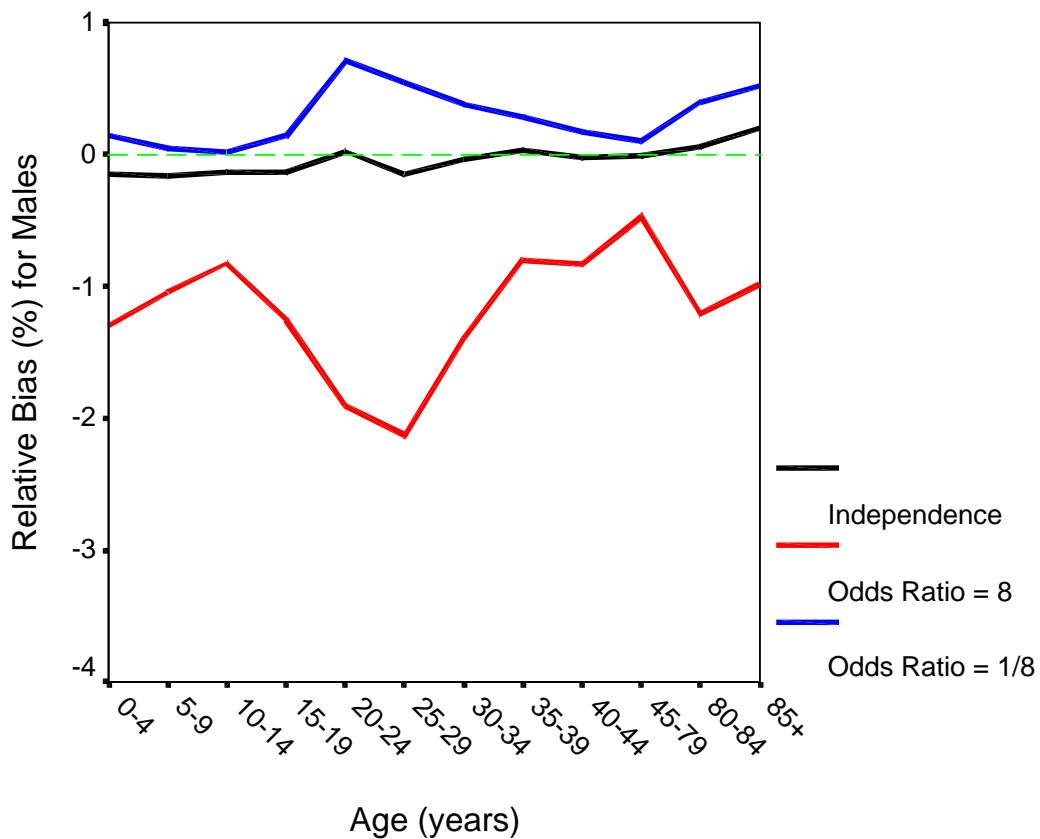
**Table 2** displays these quantities calculated from the simulations for the total population.

**Table 2: Results for the Total Population**

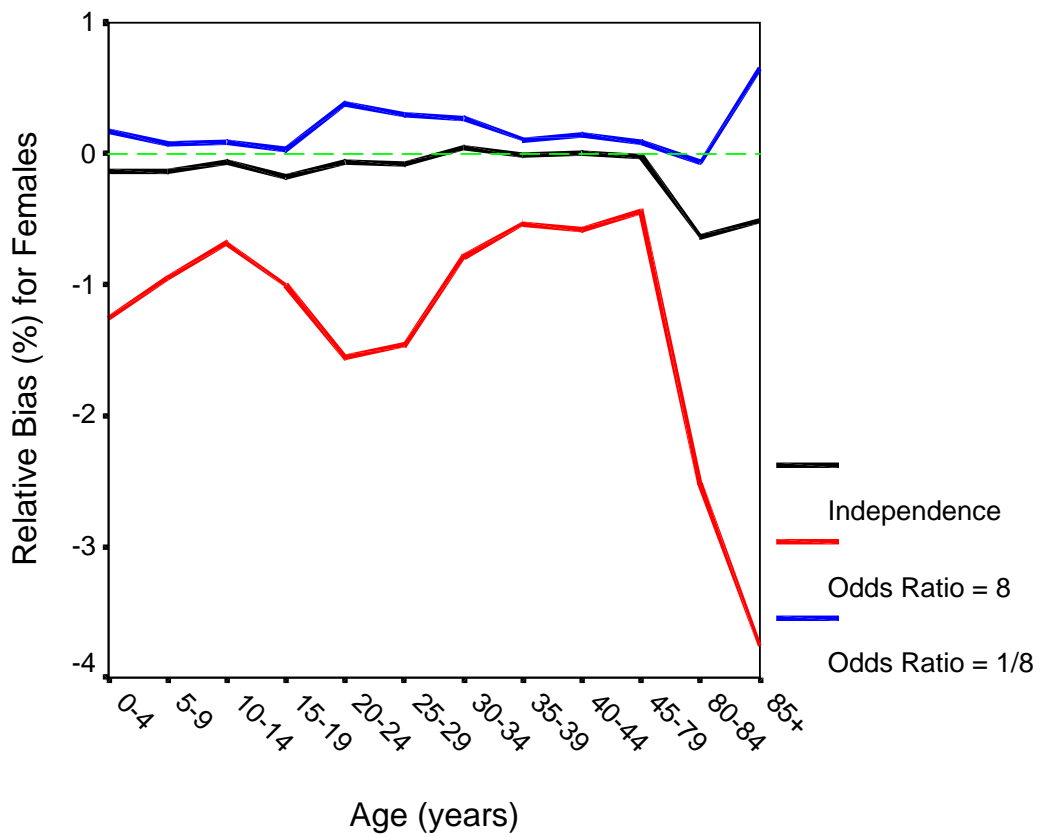
	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Independence	-0.07	0.48	-4.43
Odds Ratio = 8	-0.96	1.05	-74.32
Odds Ratio = 1/8	0.18	0.53	11.69

2.1.8 Table 2 demonstrates how the estimator behaves at the total population under dependence. The odds ratio of 8 induces negative bias, and the relative RMSE increases due to the increase in absolute bias. This is expected as an odds ratio greater than one implies the CCS is missing more of those missed by the census than it should. However, compared to the census (bias of -5 per cent) the estimator is still considerably better, and the bias is not the -2 per cent suggested in paragraph 2.1.6. This is because the calculation in paragraph 2.1.6 is based on a single DSE at the population level with fixed coverage in the census and CCS. The actual estimation procedure uses the DSE within age-sex groups at the postcode level and at this level the coverage in both the census and CCS vary. For the odds ratio less than one, the bias increases but not as dramatically as the drop for odds ratios greater than one. This is what the calculations in 2.1.3 suggest. However, as with odds ratios greater than one the result is not quite what those calculations imply due to impact of varying census and CCS response rates on the estimation strategy. The effect of this is demonstrated in **Figures 1, 2, 3 and 4**.

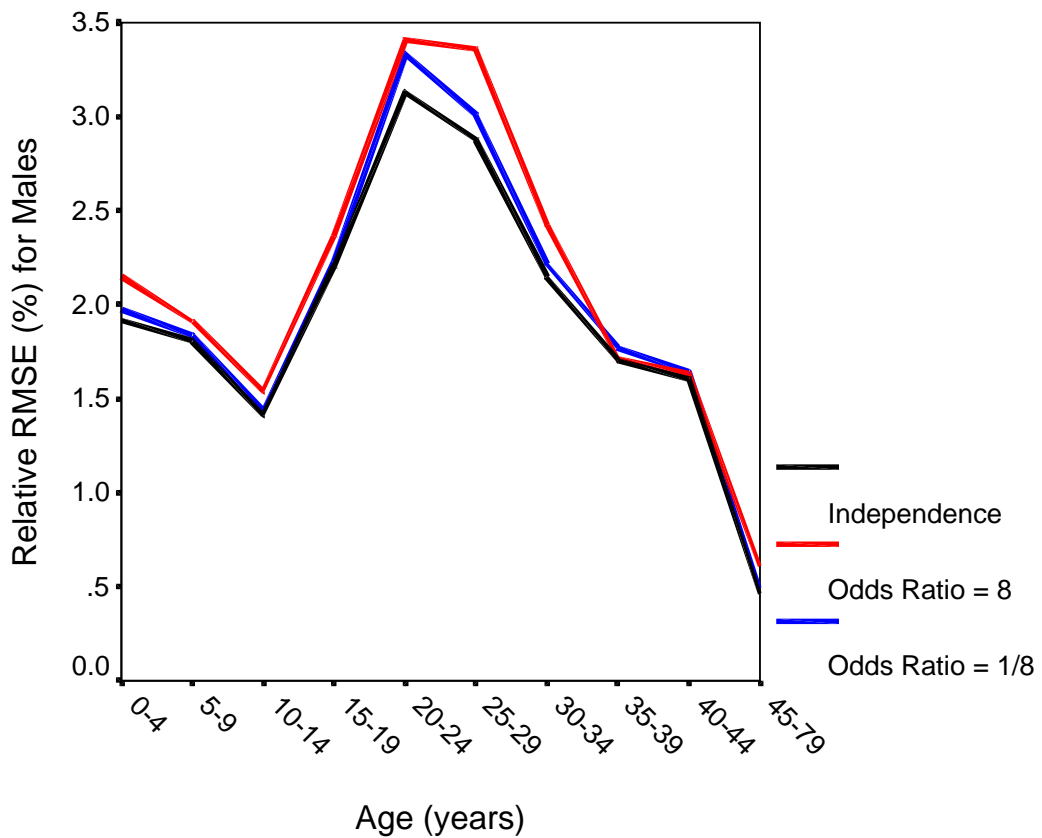
*Figure 1: Relative Bias of Male population estimates by agegroup*



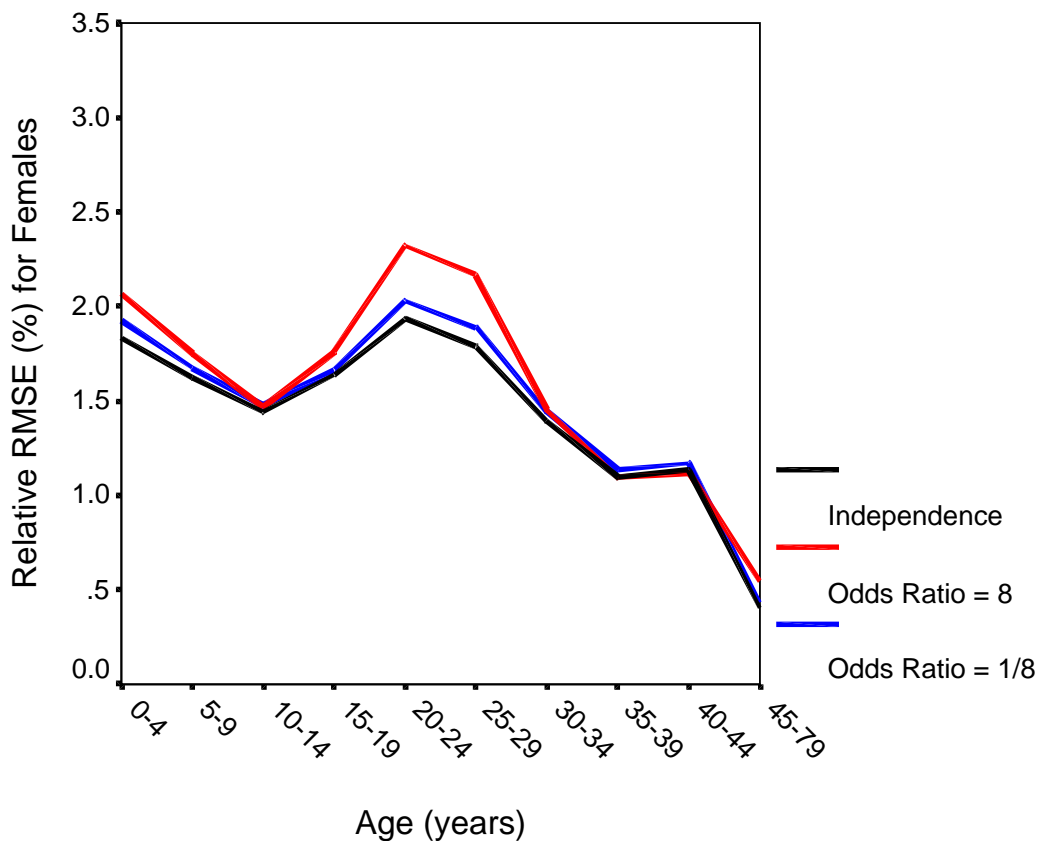
**Figure 2: Relative Bias of Female population estimates by agegroup**



**Figure 3: Relative RMSE of Male population estimates by agegroup**



**Figure 4: Relative Bias of Female population estimates by agegroup**



2.1.9 **Figures 1 and 2** demonstrate how the impact of dependence varies for different levels of coverage. For example, the age-sex group males aged 20-24 has lower census coverage and lower CCS coverage than the overall average for the simulations. This is reflected in Figure 1 by a greater impact on the bias due to dependence compared to other age-sex groups. Conversely, for females aged 45-79 Figure 2 demonstrates that the high census and CCS response rates for this group mean a much smaller gap between the different dependence scenarios. Across all age-sex groups, the bias is still much closer to zero than the unadjusted census counts which will have an average negative bias of 5%. However, the pattern in the bias is starting to reflect the census coverage with increases in the negative bias for young men and old women. For an odds ratio of 1/8, the pattern of positive bias is a reflection of this but not so pronounced. This reflects the change in CCS coverage, particularly for young men.

2.1.9 It must be noted that the simulations assume the same level of dependence everywhere – this is not likely to be particularly realistic as we would expect it to vary across area and population types. Therefore the results show the worst case scenario – the impact on the bias of the estimator would be lessened if this extreme level of dependence were to occur in just a few EDs within a particular Hard to Count category.

2.1.10 The results show that as the Census and CCS response rates fall, the estimation strategy becomes more susceptible to the effects of dependence. This reinforces the need for a good Census and CCS in the first place. It also

indicates that it is important that the CCS does well in the areas most susceptible to undercount – i.e. within the hardest to count strata.

## **Conclusions**

2.1.11 **Figures 1, 2, 3 and 4** demonstrate that dependence is important. However, for even quite extreme levels of dependence, the impact is small provided both the Census and CCS have high response rates. As the response rates fall, the estimation strategy becomes more susceptible to the effects of dependence. This can be seen by looking at how dependence effects the different age-sex groups which all have different census and CCS coverage but the same odds ratio between the two counts. Having said that, it is worth noting that in the situation of odds ratios greater than one (intuitively the most likely scenario), the estimation strategy will, in terms of bias, always do better than the unadjusted 2001 Census.

## 2.2 Sensitivity of the adjusted ratio estimator to the choice of adjustment for prediction of large census counts

### Introduction

2.2.1 The robust approach to prediction outlined in Steering Committee paper ONS(ONC(SC))00/03A is based on the philosophy of m-estimation used in robust estimation of model parameters, but extended to prediction. The approach used in m-estimation is to allow observations to have full influence over a certain range and then reduce that influence as the residual associated with an observation increases. It is then a case of choosing a sensible function that defines how this happens. There are three basic ideas:

- i) Trim observations so once residuals exceed some value they have no influence.
- ii) Decrease the influence of observations once residuals exceed some value such that eventually very large residuals mean the observation has no influence (Hampel type influence functions).
- iii) Keep the influence of observations at a constant level once residuals exceed some value (Huber type influence functions).

2.2.2 The approach presented in ONS(ONC(SC))00/03A had several components, the main part being prediction for postcodes with census counts outside the range of the CCS data. An approach with a similar ethos to (ii) was used for this component. The predicted adjustment for census counts outside the range of the CCS postcodes was reduced so that postcodes with census counts more than twice that seen in the sample postcodes have a zero adjustment. In other words the influence of increasingly extreme census counts on the estimated total underenumeration is reduced until it reaches zero.

2.2.3 It was suggested that this approach be compared to an approach with a similar ethos to (iii) such that all postcodes with census counts beyond the range of the sample data have a constant adjustment for underenumeration. In other words they all have the same influence on the estimated underenumeration regardless of the actual value of the census count. This can be thought of as a line that is parallel to the  $Y=X$  axis – the vertical distance between them is determined by the slope of the ratio line up to the point of the largest CCS postcode.

2.2.4 In general m-estimators based on influence functions like (ii) have a greater negative bias than those based on (iii). However, the advantage of this is usually a gain in the overall error (Mean Square Error) from reductions in variance due to the approach having a greater impact on the extreme values that increase variance.

## Methodology

2.2.5 The two alternatives were implemented within a set of simulations together with a standard ratio model. The simulation methodology is the same used for the previous Design Group estimation research presented in ONS(ONC(SC))00/03A. This enabled comparisons to be made between the standard ratio estimator (using the cluster level DSE), the robust ratio estimator using the approach outlined in ONS(ONC(SC))00/03A and paragraph 2.2.2 (referred to as robust ratio 1), and the robust ratio estimator using the alternative approach outlined above in paragraph 2.2.3 (referred to as robust ratio 2).

## Results

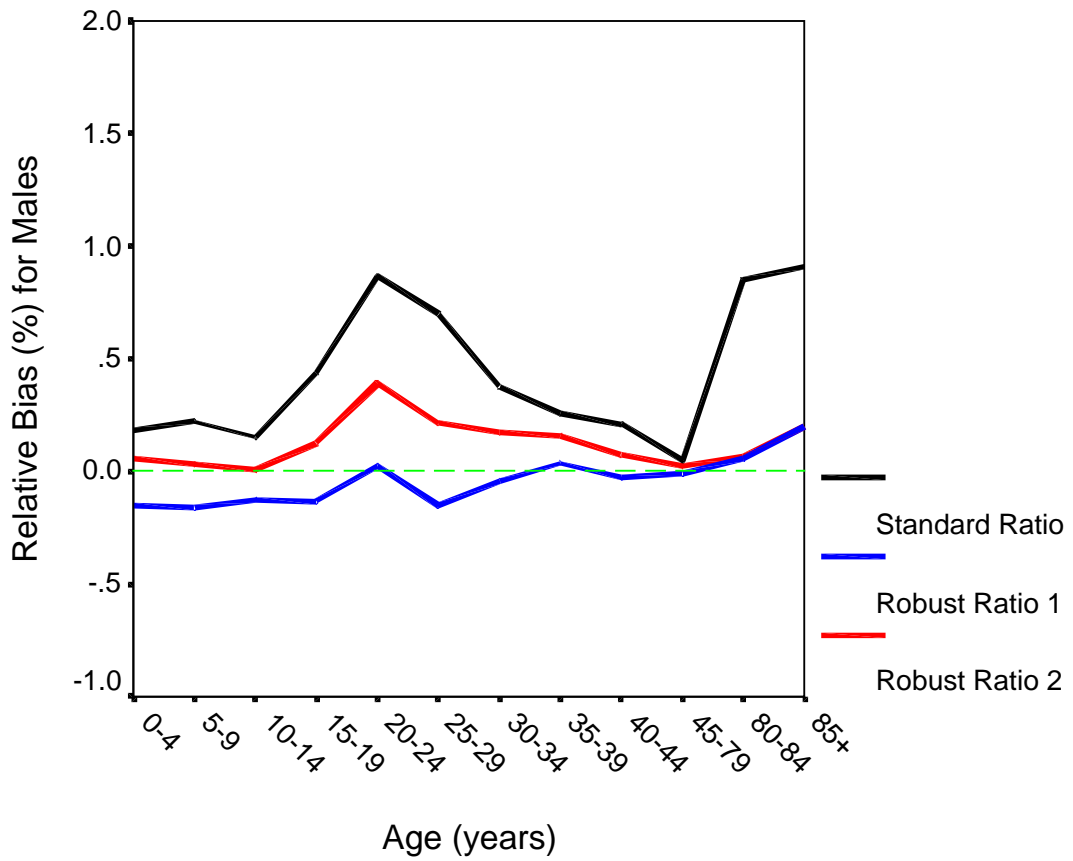
2.2.6 **Table 3** displays the Relative Bias, Relative RMSE and Z values for the Bias of the total population estimates from each of the three approaches. At the total population level there is little to choose between the alternative robust approaches, and the results are as expected. Robust ratio 1 (Hampel type approach) has a slight negative bias reflecting the correction made for predictions from extreme points. Robust ratio 2 (Huber type approach) still has a slight positive bias, although it is reduced compared to the standard ratio. This reflects the fact that extreme points still have some influence when predicting for the estimation of population totals.

**Table 3: Results for the Total Population**

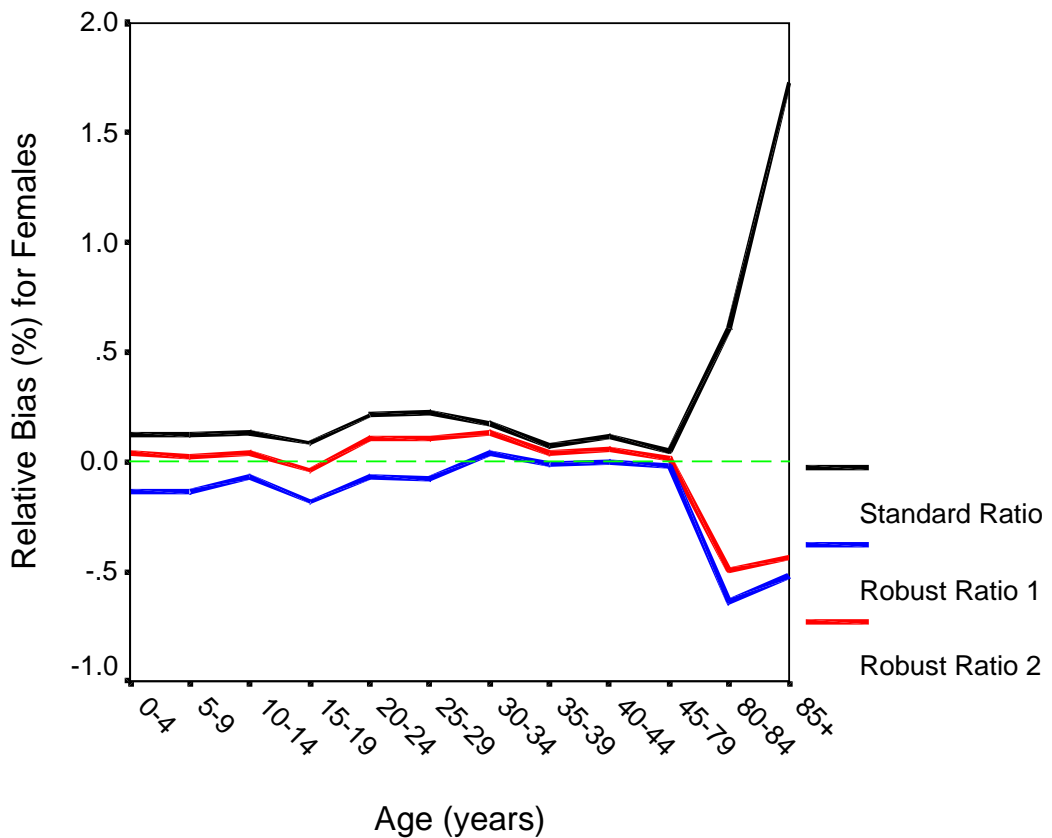
	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Standard Ratio	0.22	0.53	13.95
Robust Ratio 1	-0.07	0.48	-4.43
Robust Ratio 2	0.06	0.48	4.00

2.2.7 Further investigations of the distribution of the Bias and MSE across agegroups are displayed in **Figures 5, 6, 7 and 8** overleaf. These will indicate whether there are any differences between estimators for individual agegroups.

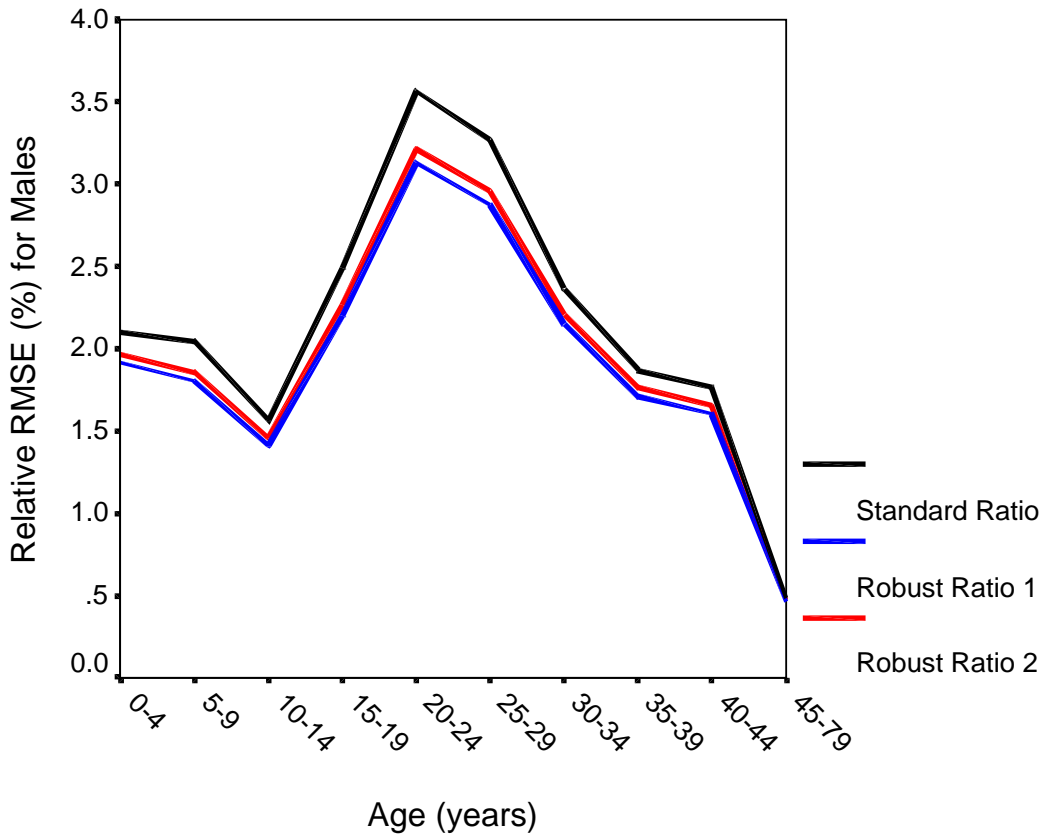
**Figure 5: Relative Bias of Male population estimates by agegroup**



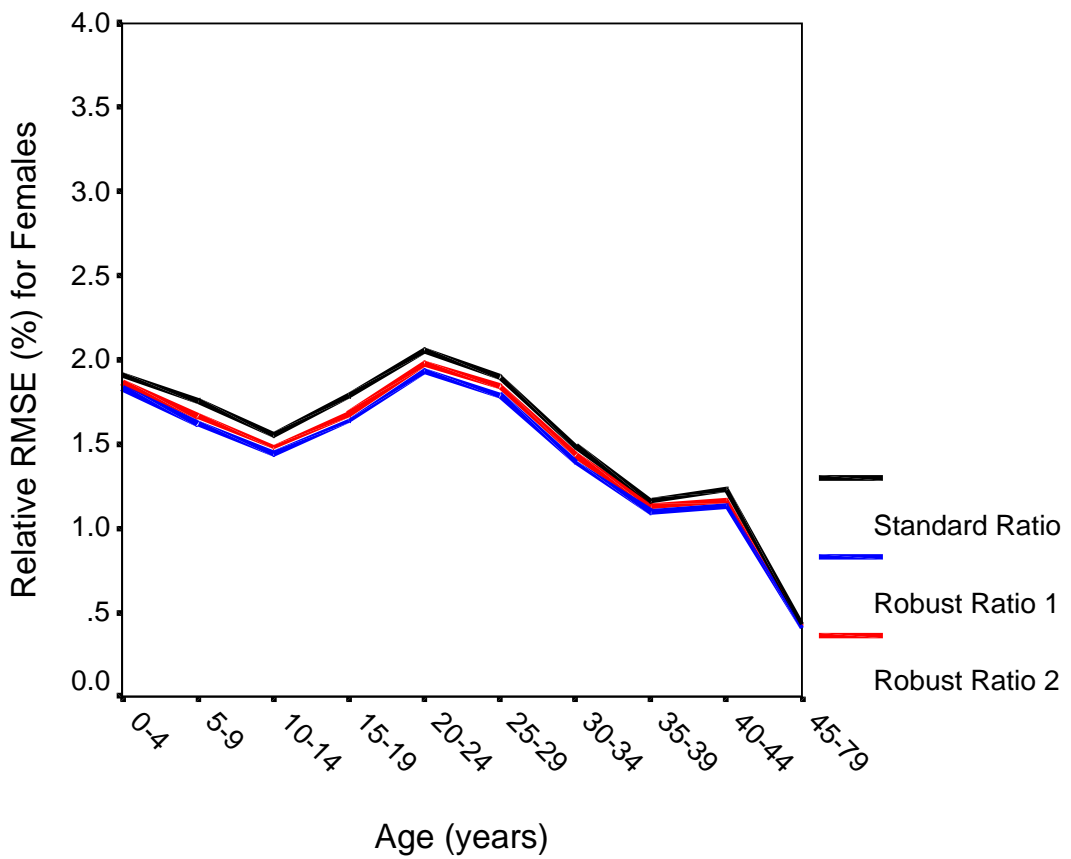
**Figure 6: Relative Bias of Female population estimates by agegroup**



**Figure 7: Relative RMSE of Male population estimates by agegroup**



**Figure 8: Relative RMSE of Female population estimates by agegroup**



2.2.7 **Figures 5 and 6** show that the bias across the age-sex groups has a similar pattern, with both robust estimators giving improvement compared to the standard ratio estimator. For males robust ratio 1 is flatter and reduces the impact of being a young male. Conversely, for females robust ratio 2 does particularly well at the younger ages. **Figures 7 and 8** confirm that robust ratio 1 does fractionally better in terms of RRMSE across all age-sex groups, as m-estimator theory suggests it should. However, the gain is only marginal.

## **Conclusions**

2.2.8 Based on the simulation results there is little to choose between the two estimators. Robust ratio 2 has intuitive appeal as it always makes some adjustment for underenumeration. Conversely, it can be argued that robust ratio 1 has better statistical properties in terms of relative RMSE. However, this makes little impact in the simulation study and any gain from adopting robust ratio 1 is marginal. Therefore, the simulations suggest no real efficiency gain from choosing robust ratio 1. This approach is also likely to have less intuitive appeal to the user community.

2.2.9 **It is therefore recommended that as the simulations do not conclusively favour either approach the revised strategy (robust ratio 2) for dealing with prediction outside the sample range is the preferred approach due to its intuitive appeal.**

## 2.3 The method of variance estimation

### Introduction

2.3.1 Estimation of variances for the Design Group estimates of the population are essential to allow quality assurance with other external estimates of the population in 2001, such as those produced nationally using demographic methods. Work in Brown *et al* (1999) used the ultimate cluster variance estimator which has a general form for the variance of an estimator  $\hat{\theta}$  given by:

$$\text{Var}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{g=1}^n (\hat{\theta}_g - \hat{\theta})^2 \quad (1)$$

where  $n$  is the number of PSUs in the sample, and  $\hat{\theta}_g$  is an estimator based only on the data from PSU  $g$ .

2.3.2 Another common, and related, variance estimator is the jackknife estimator which has a general form for the variance of an estimator  $\hat{\theta}$  given by

$$\text{Var}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{g=1}^n (\{n\hat{\theta} - (n-1)\hat{\theta}^{(g)}\} - \hat{\theta})^2 \quad (2)$$

where  $n$  is the number of PSUs and  $\hat{\theta}^{(g)}$  is an estimate based on all the data excluding PSU  $g$ .

2.3.3 In both (1) and (2) a finite population correction can be included at the front but for large population sizes it will make very little difference. They can also be generalised to allow the estimation of a covariance between two different estimators.

2.3.4 These techniques were applied to the simulation data, within HtC stratum, for the robust ratio 1, the estimator in Steering Committee paper ONC(SC)00/03A. It was necessary to not only estimate the variances for the age-sex estimates but the complete variance-covariance matrix to allow the computation of a variance for the estimate of the total population.

### Results

2.3.5 As this is a simulation the empirical variance of the estimator across the simulation approximates the true variance of the estimator. The empirical variance for the estimate of the total population from the simulation is 4,478,323. **Table 4** below gives the mean value and coverage for the two variance estimators for the total population

**Table 4: Performance of variance estimators for the total population**

Type of Estimator	Mean Value	Coverage <sup>1</sup>
Ultimate cluster	7,936,568	0.963
Jackknife	5,065,500	0.944

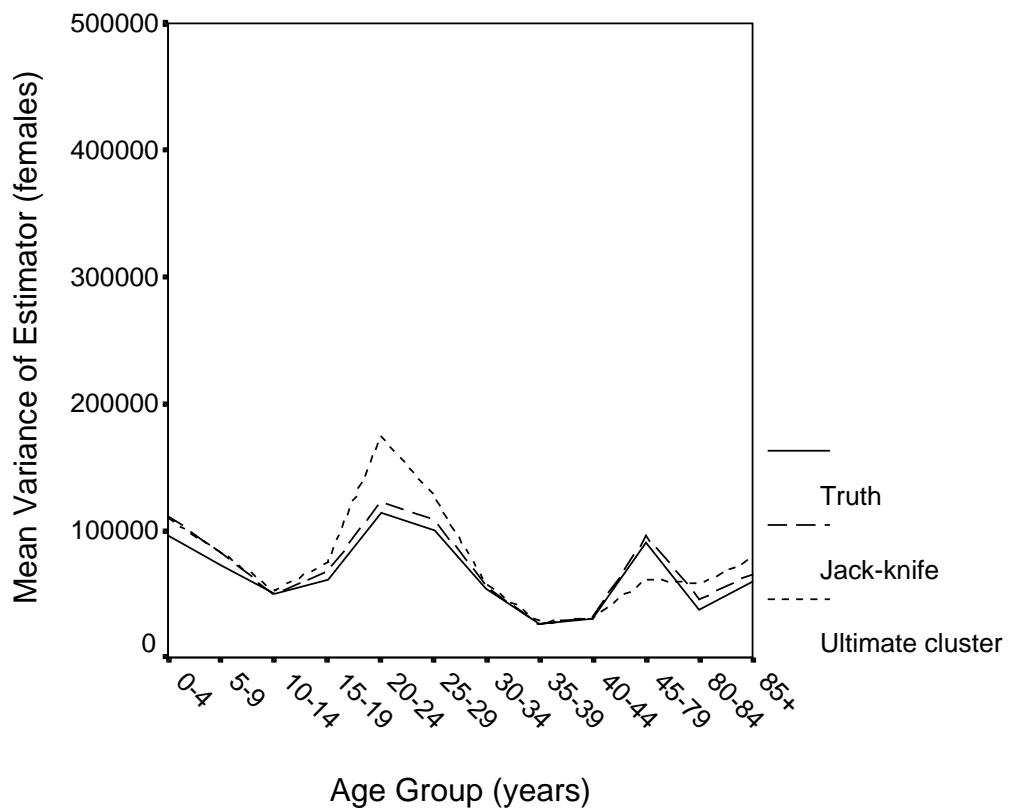
<sup>1</sup> Based on estimated 95% confidence intervals

2.3.6 Comparing the mean values in **Table 4** with the empirical variance of 4,478,323 demonstrates that both estimators are conservative (positively biased), and this is particularly true for the ultimate cluster variance estimator. However, in both cases the coverage for a 95% confidence interval is approximately correct, one being slightly under and one slightly over. **Figures 9, 10, 11 and 12** present the same results but for the individual age-sex estimates. The truth is the empirical variance from the simulation for the individual age-sex groups.

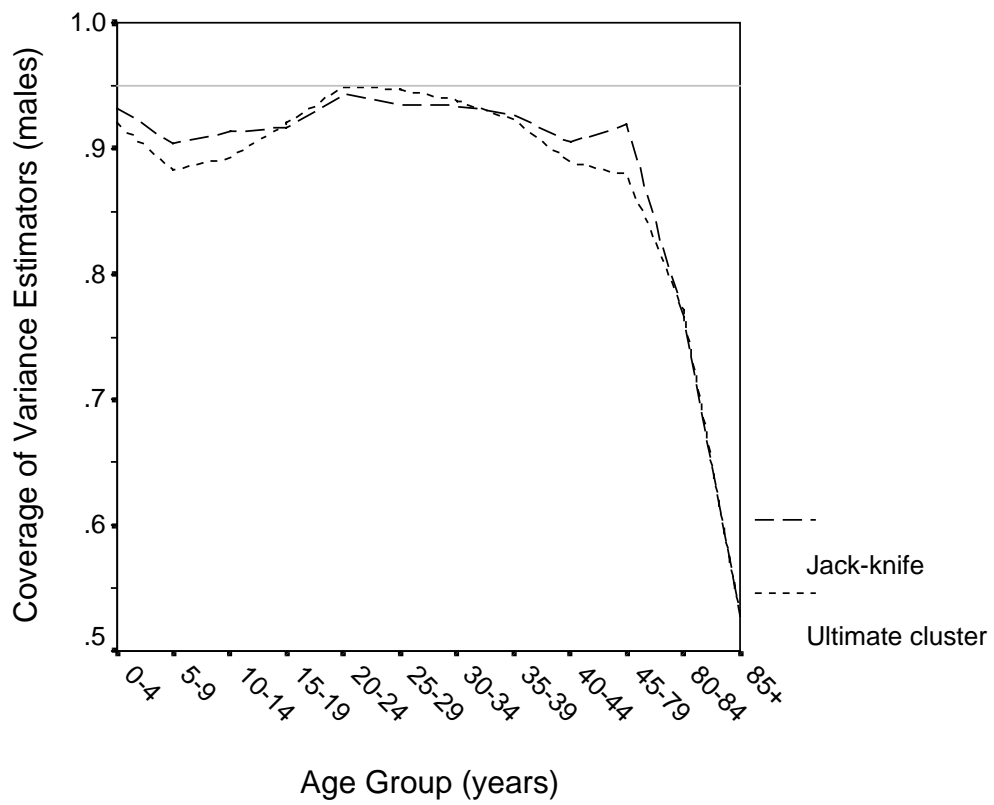
**Figure 9: Mean Variance of Estimator for males by agegroup**



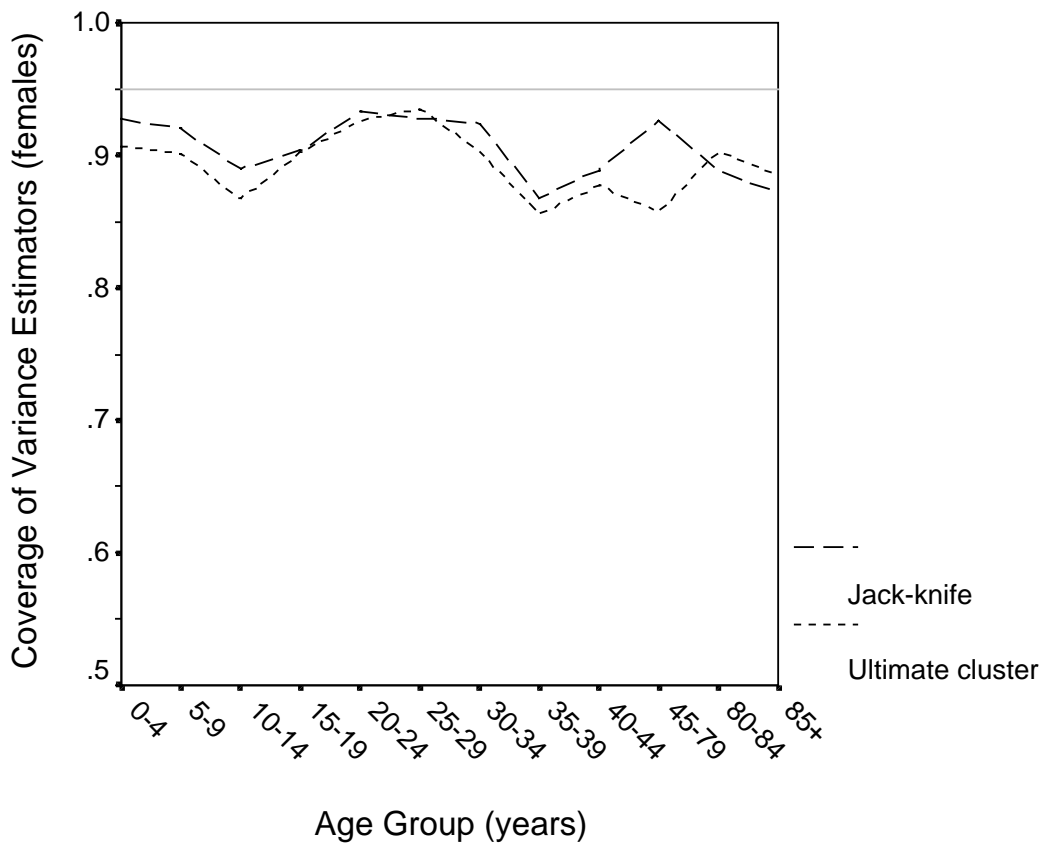
**Figure 10: Mean Variance of Estimator for females by agegroup**



**Figure 11: Coverage of Variance Estimators for males by agegroup**



**Figure 12: Coverage of Variance Estimators for females by agegroup**



2.3.7 **Figures 9 and 10** show that the jackknife variance estimator tracks the true variance more closely. This is particularly true for young adult males and, to a lesser extent, the same agegroups for females. Both variance estimators rely on variation between estimates based on part of the sample. In particular, the ultimate cluster variance estimator looks at variation between estimates based on single PSUs. It is intuitive that the complex robust ratio estimator will be less stable when estimated with small amounts of data, especially for agegroups where there is considerable and more variable underenumeration. In certain cases, such as males aged 20-24, this lack of stability is over-stating the variability of the estimator. **Figures 11 and 12** demonstrate that both estimators fall slightly short of achieving 95 per cent coverage for their estimated confidence intervals although across all age-sex groups the jackknife is doing slightly better.

### **Conclusions**

- 2.3.8 Although the two approaches are related, the ultimate cluster variance estimator, which relies on estimates based on single PSUs, will be more unstable. Both estimators have coverage problems in that they do not give 95% coverage for 95% confidence intervals. This is a particular problem for the 80-84 males and the 85+ males. This is caused by variance estimates of zero when the CCS fails to find any extra people over the census in the CCS sampled postcodes. Such situations will occasionally arise in 2001 and further work is needed to specify a strategy for collapsing age-sex groups to allow variances to be estimated.
- 2.3.9 **It is recommended that a jackknife variance estimator is used for the estimation variances associated with the ONC Design Group population estimates.**

### **3. LOCAL AUTHORITY DISTRICT ESTIMATION**

#### **3.1 The agegroup categories used in the LAD estimation models**

##### **Introduction**

- 3.1.1 Population estimates of five year age-sex groups are produced directly within the Design Groups. However, because of the small sample sizes involved for the LAD estimation the models evaluated in Paper ONC(SC)00/03B used a collapsed set of age groupings, with the exception of the simple synthetic estimator (since this uses the already derived five year age/sex Design Group estimates). By collapsing together some of the categories, the assumption was made that the underenumeration does not vary for the categories that have been collapsed. The reduced groups were used to fit the model but predictions were calculated separately for the uncollapsed categories.
- 3.1.2 This research examines whether the collapsed agegroups used in paper ONC(SC)00/03B provided estimates of a higher precision for all methods than if the agegroups had not been collapsed.

##### **Methodology**

- 3.1.3 To examine the effect of the agegroups used in the models, a simulation study was undertaken to compare the collapsed agegroup models with an uncollapsed version. The simulation methodology used was identical to paper ONC(SC)00/03B. The sixteen collapsed categories used previously were:

- 0-4 year olds
- 5-14 year olds
- 15-19 year old males
- 15-19 year old females
- 20-24 year old males
- 25-29 year old males
- 20-29 year old females
- 30-34 year old males
- 35-39 year old males
- 30-39 year old females
- 40-44 year olds
- 45-59 year olds
- 60-69 year olds
- 70-79 year olds
- 80+ males
- 80+ females

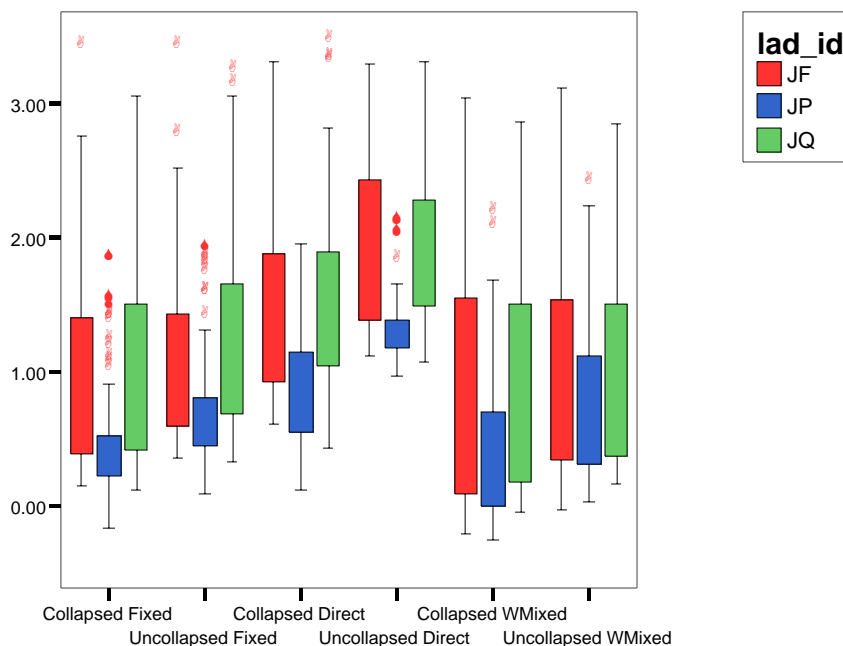
- 3.1.4 These collapsed categories were used to fit all the models (with the exception of the simple synthetic model), and estimates were calculated for the 36 age-sex categories through use of the fitted collapsed model. A comparison will be made between models fitted using:
- a) the set of 16 collapsed groups; and;
  - b) the 36 uncollapsed age-sex groups (as used in the Design Group models);

3.1.5 The approaches compared were the Direct, Synthetic, LAD adjusted synthetic and the Weighted Mixed models. Simulations were carried out for a selection of five of the Design Groups used in the previous research. The groups used were Hampshire A versions 1 and 3, Hampshire B versions 3 and 4 and Hampshire C version 1 (see **Annex A** for details of the makeup and coverage rates for these areas). This selection of Design Groups were chosen as they provided the key comparisons between the approaches in the previous research.

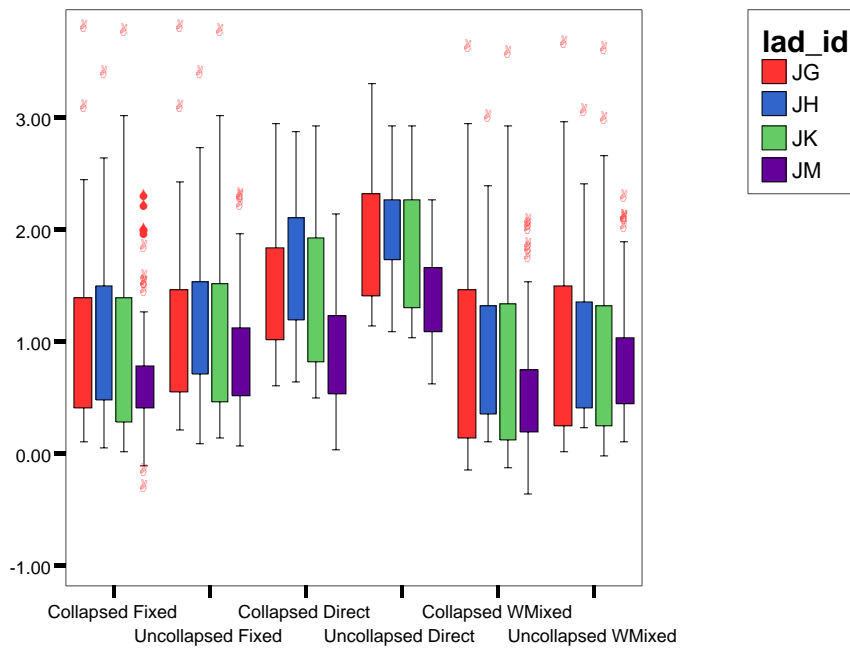
## Results

3.1.6 The key measure of performance of the different approaches is the Relative Root Mean Squared Error (RRMSE) of the total population estimates for the five year age-sex group within each LAD. These actual error values can be calculated since the true value of these populations are known. The great benefit of using this particular measure of error is that it contains both a variance and bias component – and therefore our evaluation can take both into account. The mean of these errors across all 1000 simulations for each subgroup are used in the analysis. As these values are extremely skewed the natural logarithm (i.e.  $\text{Log}_e$ ) of the Mean RRMSEs are utilised for presentational purposes. A description of the RRMSE calculations can be found in Annex B.

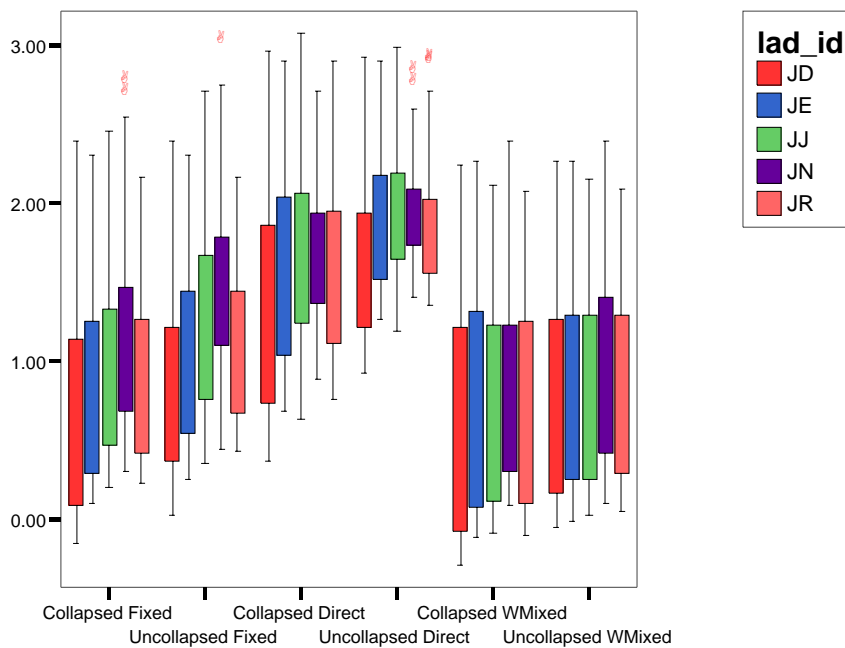
**Figure 13 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire Group A version 3 (JF is Eastleigh, JP is Southampton and JQ is Test Valley)**



**Figure 14 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire Group B version 3 (JG is Fareham, JH is Gosport, JK is Havant and JM is Portsmouth)**



**Figure 15 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire Group C version 1 (JD is Basingstoke, JE is East Hampshire, JJ is Hart, JN is Rushmoor and JR is Winchester)**



- 3.1.7 **Figures 13, 14 and 15** display the distribution of the age-sex group by hard to count index estimation errors for each of the LADs in the Design Groups. The results for both the collapsed and uncollapsed models are shown to allow a comparison. They all show without exception that the collapsed models are more efficient than the uncollapsed models. As an example, when comparing the paired collapsed and uncollapsed approaches in **Figure 14**, the purple boxplots which represent the Portsmouth LAD are clearly lower on the precision scale for each of the collapsed approaches.
- 3.1.8 This is a good indication that the assumption that the undercount was similar within the collapsed groups was not violated. For instance, the coverage for males aged 45 to 59 was similar to females aged 45 to 59.
- 3.1.9 These particular results arise from the way in which the simulations were carried out. Because we were able to examine the actual simulated census coverage within each agegroup, the collapsed groupings were chosen to ensure that the assumption of similar coverage was not violated. However, this will not be possible in 2001. Therefore, consideration must be given to whether agegroups should be collapsed for 2001.

## **Conclusions**

- 3.1.10 Previous research evaluated a number of models to determine the best methodology for the production of LAD population estimates. The results of this research have indicated that for all of the approaches examined, the use of a collapsed set of agegroups within the models provided improved precision when compared to models that did not collapse agegroups. **Therefore the results of the previous research still hold. Furthermore, it is recommended that the collapsing of agegroups should be considered further for the ONC Local Authority District estimation.**

## 3.2 Composite LAD estimators

### Introduction

- 3.2.1 The results presented in paper ONC(SC)00/03B show that for the majority of simulations carried out, the simple apportionment estimator provides population estimates with the greatest precision. However, when the assumption of no difference in the underlying LAD coverage is seriously violated, the model fails resulting in biased predictions. It was therefore concluded that the approach could not be applied with confidence.
- 3.2.2. The robustness of the chosen model is the most important aspect as there are likely to be many different conditions across the 101 Design Groups in 2001. It was therefore recommended that the LAD specific fixed estimator be applied within the 2001 ONC estimation strategy, as this approach has less potential for producing biased estimates.
- 3.2.3 However, while this estimator is more robust, it is not as efficient as the simple synthetic estimator. It would therefore be sensible to explore whether a composite estimator of the form (3) would inherit the efficiency and robustness properties of the two alternatives – in other words provide the best of both worlds. Composite estimators are discussed in detail in Ghosh and Rao (1994).

$$\hat{T}_{adl}^C = (1 - w_{adl})\hat{T}_{1adl} + w_{adl}\hat{T}_{2adl} \quad (3)$$

Where:

$\hat{T}_{1adl}$  is the population estimate from the simple synthetic estimator

$\hat{T}_{2adl}$  is the population estimate from the LAD specific fixed estimator

$w_{adl}$  are appropriately chosen weights

- 3.2.4 This research examines whether the potential gains from such an approach can be made. It is also important to consider how such an approach would impact on the estimation of variances. In this case, it would make the variance estimation process extremely complex and therefore the efficiency gains would have to be significant to warrant the adoption of a composite estimator.

### Methodology

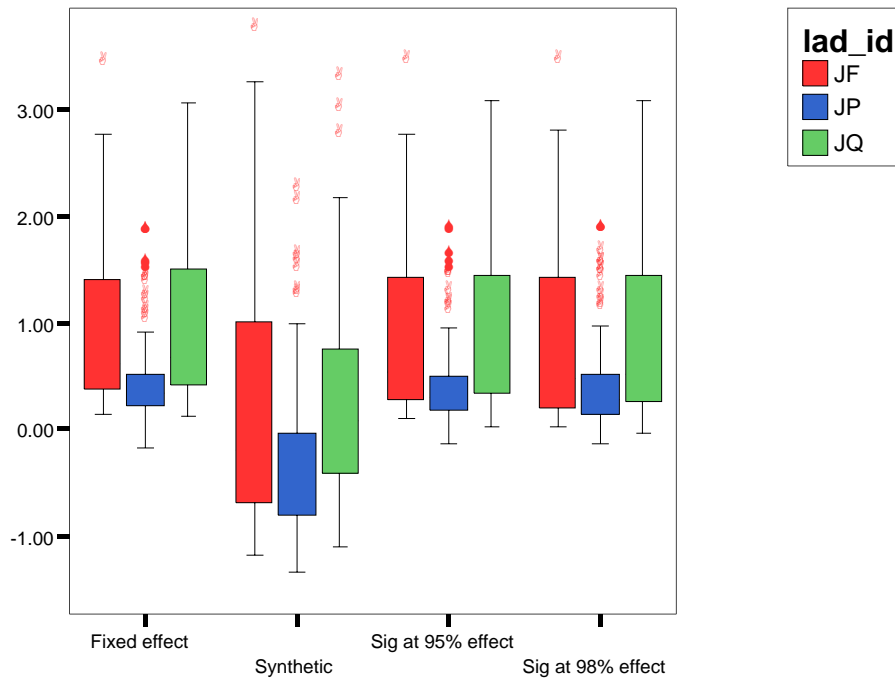
- 3.2.5 The main issue in the development of a composite estimator (3) is the formulation of the weights  $w_{adl}$ . Simple methods include choosing the weights such that the Mean Square Error of the total estimator is minimised, or basing weights upon the sample size (see Ghosh and Rao (1994)). However, noting that the LAD specific fixed estimator is an extension of the simple synthetic estimator (ignoring the collapsed agegroups) then it might be appropriate and simpler to evaluate the inclusion of the LAD specific effect.

- 3.2.6 This could be achieved through evaluating whether the LAD effect is predictively useful in the model (in other words, carry out a hypothesis test that the parameter estimates are significantly different from zero). If it is useful we set  $w_{adj}=1$  and the LAD adjusted model is used. If there is no evidence of a significant LAD effect,  $w_{adj}=0$  and a synthetic apportionment approach is used.
- 3.2.7 To carry out the hypothesis test, the LAD adjusted model is fitted and the resulting F statistics for the LAD specific parameters are examined. For a two tailed test of size  $\alpha=0.05$ , if any of the F statistics are  $>1.96$  then we can conclude that there is a significant difference between the LADs and the weight  $w_{adj}=1$ .
- 3.2.8 In order to evaluate the composite model proposed above, it was implemented within a set of simulations in order to compare its performance with the simple synthetic model and LAD adjusted synthetic model. Two versions were examined – one that used a test of size  $\alpha=0.05$  and one that used a test of size  $\alpha=0.02$ . The simulations were carried out on the same basis as those in paper ONC(SC)00/03B.

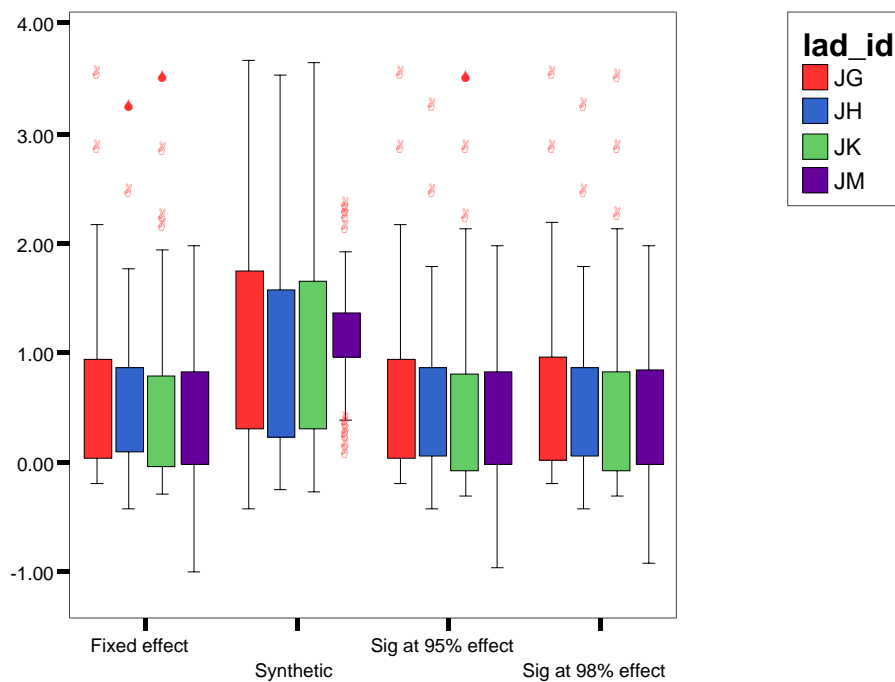
## Results

- 3.2.9 The key measure of performance used is the Relative Root Mean Squared Error (RRMSE) of the total population estimates for the five year age-sex group within each LAD. These actual error values can be calculated since the true value of these populations are known. This measure of error contains both a variance and bias component – and therefore our evaluation can take both into account. The mean of these errors across all 1000 simulations for each subgroup are used in the analysis. As these values are extremely skewed the natural logarithm (i.e.  $\text{Log}_e$ ) of the Mean RRMSEs are utilised for presentational purposes. A description of the RRMSE calculations can be found in Annex B.

**Figure 16 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire Group A version 1 (JF is Eastleigh, JP is Southampton and JQ is Test Valley)**



**Figure 17 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire Group B version 3 (JG is Fareham, JH is Gosport, JK is Havant and JM is Portsmouth)**



- 3.2.10 **Figures 16 and 17** show the distribution of the age-sex group by hard to count index estimation errors for each of the LADs in the Design Groups. They both show that there is not a great deal of difference between the fixed effect model and the two composite estimators. For **Figure 17** where there should be big differences in the LAD undercount, this is as we would expect. However, we would expect to be making efficiency gains where there are no differences as in **Figure 16**.
- 3.2.11 The reason for this may be that the test for an LAD effect is detecting very small differences – and so even in the cases where the LAD undercount are not that different the composite estimator is the equivalent of the Fixed LAD estimator. A solution could be to alter the size of the test to push up the significance levels. However, the evidence here indicates that the size would have to be pushed up quite a lot to make the expected gains. This carries some risk with it, since it may affect the robustness of the estimator.

### **Conclusions**

- 3.2.12 The results clearly show that the composite estimators are not significantly more efficient than the LAD specific fixed estimator. Even though gains are made, there are added complications of variance estimation when using such estimators. There is therefore not enough evidence to suggest altering our original strategy for estimating the LAD populations.
- 3.2.13 **It is therefore concluded that composite estimators do not provide enough gains in efficiency to warrant their use.**

### 3.3 Inner London LAD Estimation

#### Introduction

- 3.3.1 At the February 2000 meeting of the ONC Steering Committee, members expressed a wish for the proposed small area estimation strategy to be tested using data from an area of inner London. Since the 1991 undercount in London is believed to be of a different nature to other areas of the country, this will provide more information on the robustness of the recommended strategy.
- 3.3.2 Following the further work outlined in this paper, 1991 Census data for Inner London were obtained. A set of data suitable for simulation was developed using an identical method to that used to derive the version 1 Design Groups in the previous studies (see paper ONS(ONC(SC))00/03B). This essentially uses the estimating with confidence coverage factors to drive the underenumeration.
- 3.3.3 A Design Group was selected consisting of Hackney, Islington and Tower Hamlets. **Table 5** displays the population and mean simulation census coverage of the LADs within the Design Group. Note that the overall coverage levels are not significantly different, although they are relatively low. Based on the previous simulation work, we can therefore expect the synthetic model to provide the best results.

*Table 5 – Inner London Design Group.*

LAD	Simulation resident population	Mean Census coverage (percentage to 2 d.p.)
Hackney	165,500	94.17%
Islington	154,000	93.73%
Tower Hamlets	153,500	94.51%

#### Methodology

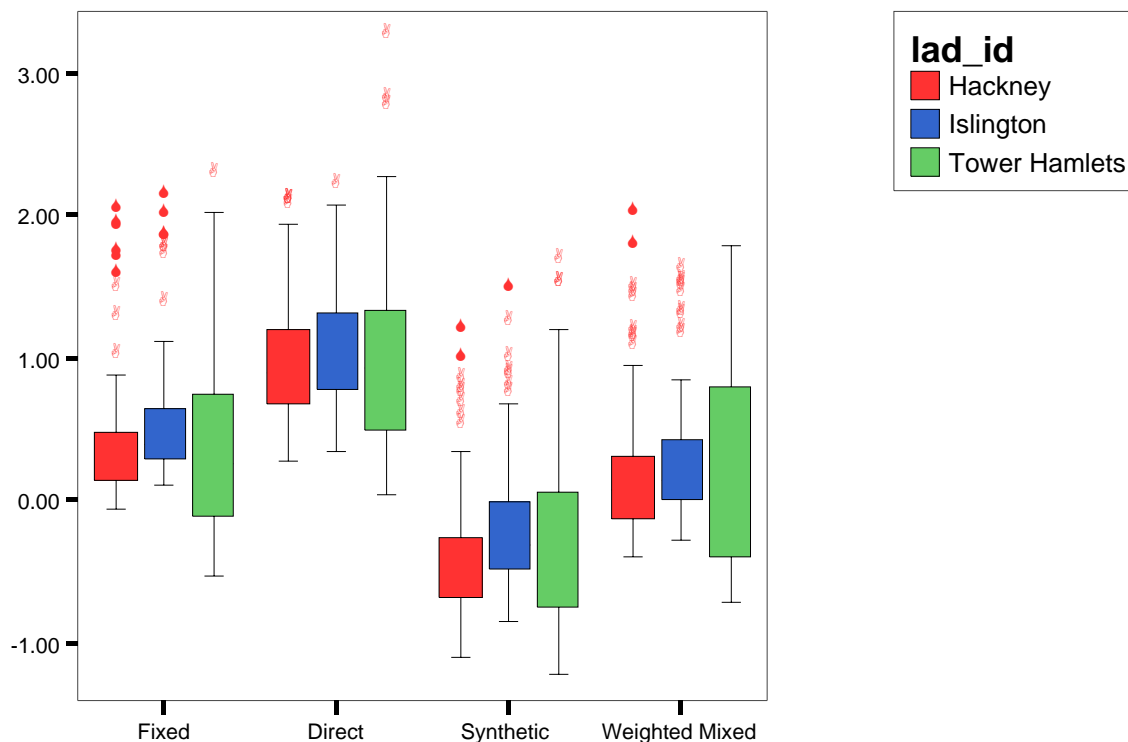
- 3.3.4 The approaches that were evaluated in the previous work were implemented for the Design Group to allow comparisons to be drawn. The following labels are used to identify the different approaches:
- a) direct ratio estimator (Approach 1) – this is labelled as **Direct**;
  - b) simple synthetic estimator (Approach 2) – this is labelled as **Synthetic**;
  - c) constrained fixed effects (Approach 3) – this is labelled as **Fixed**;
  - d) weighted mixed model (Approach 4) – this is labelled as **Weighted Mixed**.
- 3.3.5 The key measure of performance used is the Relative Root Mean Squared Error (RRMSE) of the total population estimates for the five year age-sex group within each LAD. As these values are extremely skewed the natural

logarithm (i.e.  $\text{Log}_e$ ) of the Mean RRMSEs are utilised for presentational purposes. A description of the RRMSE calculations can be found in Annex B.

## Results

3.3.6 **Figure 18** shows the distribution of the  $\text{Log}_e$  RRMSEs across the LADs for each approach. It clearly indicates that as expected the synthetic approach provides the greatest precision. Similarly, the performance of the other approaches is comparable with previous research. This indicates that the conclusions presented in paper ONS(ONC(SC))00/03B are likely to hold in these type of areas if the coverage patterns are similar.

**Figure 18: Distribution of Ln RRMSEs for each Local Authority District in Inner London Design Group**



## Conclusion

3.3.7 The results presented above coincide with the conclusions arising from the previous research. Because the LAD coverage patterns were quite similar within the Design Group, the synthetic estimator performed the best. However, it must be noted that coverage patterns are likely to be very different in the 2001 Census and therefore the choice of a more robust estimator is a sensible choice. Therefore, **the recommendation to use the fixed LAD effect model still holds.**

**References:**

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999)<sup>1</sup> A methodological strategy for a One Number Census. *Journal of the Royal Statistical Society A* **162**, 247-267.

Ghosh and Rao (1994), Small Area Estimation: An Appraisal, *Statistical Science*, Volume 9, No 1, pp55-93.

ONS(ONC(SC))00/03A – Estimation strategy for Design Group Estimates by age and sex from the Census Coverage Survey.

ONS(ONC(SC))00/03B – One Number Census Local Authority Estimation.

ONS(ONC(SC))00/10 – Design Groups for 2001

ONS(ONC(SC))00/13 – CCS Methodology

## ANNEX A – SIMULATION COVERAGES

*Table A1: Mean Simulation Coverage by LAD for each Design Group variant*

Local Authority District	Design Group	Mean Census LAD coverage (percentage to 2 d.p.) by simulation			
		Version 1	Version 2	Version 3	Version 4
Eastleigh	Hampshire A	95.87	95.49	95.87	N/A
Southampton	Hampshire A	93.99	92.19	92.19	N/A
Test Valley	Hampshire A	95.62	94.68	95.62	N/A
Fareham	Hampshire B	96.06	95.66	96.06	98.37
Gosport	Hampshire B	95.01	93.61	95.01	97.92
Havant	Hampshire B	95.97	95.60	95.97	98.32
Portsmouth	Hampshire B	93.08	90.52	90.52	90.49
Basingstoke	Hampshire C	95.99	95.53	95.53	N/A
E. Hampshire	Hampshire C	95.87	95.27	95.87	N/A
Hart	Hampshire C	95.74	94.81	95.74	N/A
Rushmoor	Hampshire C	93.46	91.01	93.46	N/A
Winchester	Hampshire C	95.53	94.67	95.53	N/A

## ANNEX B – Description of terms used

### a) Relative Root Mean Square Error

For a given population quantity such as the total  $T$  with estimator  $\hat{T}$ , one can measure the mean accuracy of the estimator using the relative root mean square error (RRMSE) defined as:

$$\text{RRMSE}(\hat{T}) = \frac{1}{T} \left\{ \sqrt{\frac{\sum (\hat{T} - T)^2}{n}} \right\} \cdot 100$$

where the summation is carried out over all the  $n$  observations of  $\hat{T}$ .

The RRMSE is a measure of the mean level of variability for a population total, relative to the population total being estimated.

### b) Mean Square Error

For a given population quantity such as the total  $T$  with estimator  $\hat{T}$ , one can measure the mean accuracy of the estimator using the mean square error (MSE) defined as:

$$\text{MSE}(\hat{T}) = \frac{\sum (\hat{T} - T)^2}{n}$$

where the summation is carried out over all the  $n$  observations of  $\hat{T}$ .

The MSE is a measure of the mean level of variability for a population total. The MSE includes a measure of the bias for the population estimate.