

**ONE NUMBER CENSUS STEERING COMMITTEE****Estimation Strategy for Design Group Estimates by Age and Sex from the Census Coverage Survey**

1. This paper reports the research towards developing a methodology for estimating the design group populations. These are the key estimates in the One Number Census process, as they add-up to give the national estimates and are control totals for the local authority district estimates. Several strategies are examined that combine dual system estimation with different population estimation models including ratio and regression models.
2. An extensive series of simulations have been used to evaluate the different strategies both in terms of robustness and efficiency. The results show that the ratio model, with modifications, combined with postcode level dual system estimation constrained to the cluster level is both efficient and robust.
3. **Members of the Steering Committee are asked to:**
  - a) **Note the results presented in the paper;**
  - b) **Agree the recommendation that a robustified ratio estimator using dual system estimation at the postcode level with a cluster constraint be the strategy for design group estimation; and;**
  - c) **Provide comments at the meeting or in writing by 23<sup>rd</sup> February 2000.**

**James Brown**  
**Department of Social Statistics**  
**University of Southampton**  
**Southampton SO17 1BJ**

# **Estimation Strategy for Design Group Estimates by Age and Sex from the Census Coverage Survey**

## **Executive Summary**

### **Introduction**

The One Number Census (ONC) project aims to estimate the level of underenumeration of both households and individuals in the 2001 Census. The postcode based Census Coverage Survey (CCS) has been designed to provide the data that will facilitate the estimation of census coverage by age and sex. The design is described in ONS(ONC(SC))00/01.

Dual System Estimation (DSE) methodology will be used to combine 2001 Census and CCS counts to estimate the true population in the sampled CCS postcodes. Generalisation of these DSE counts from the sampled areas to the whole population can be carried out using a variety of survey estimation methods.

The ONC estimation strategy has two phases. Phase one is the estimation of underenumeration by age and sex for approximately one hundred 'design groups' in England and Wales. Each design group is an aggregation of a number of whole Local Authority Districts (LADs) and is the level at which the CCS is designed to provide direct estimates of an acceptable precision. Phase two is the allocation of the estimated underenumeration to the individual LADs, the methodology for which is contained in ONS(ONC(SC))00/03B.

This paper focuses on the methodology which will be adopted for the production of design group population estimates as part of the ONC estimation strategy. A number of options are examined that combine dual system estimation with different population estimation models including ratio and regression models. A series of simulations are used to evaluate the different strategies both in terms of robustness and efficiency.

### **Dual System Estimation (DSE)**

Dual System Estimation is an established technique for estimating the true population size when you have two 'independent' attempts to count that population. In this case the population is an age-sex group in a postcode and the two attempts are the 2001 Census and CCS. The aim is to produce a set of adjusted counts for the areas sampled by the CCS that represent the true population accounting for individuals missed by both the 2001 Census and the CCS.

### ***Options***

The paper considers the calculation of dual system estimates at different levels of aggregation. There are advantages to increasing the level of aggregation, as the DSE becomes more stable and the relative variance of the estimator decreases. However, the assumptions of homogeneity of capture probabilities and independence between the 2001 Census and the CCS become increasingly unrealistic as the level of aggregation increases.

### ***Results***

The simulations clearly show that the individual postcode is an unsatisfactory level at which to apply the Dual System Estimator. However, the cluster of postcodes selected from each 1991 Enumeration District (ED) level offers a good compromise between stability and the dual system assumptions. However, it is

attractive for both phases of estimation to have a set of adjusted postcode counts and therefore a postcode level model. A compromise is made by constraining the postcode level estimates so that they sum to the cluster level estimates.

## **Design group estimation**

The estimation of the population at the design group level is key to enabling the ONC aims to be achieved. The CCS has been designed so that the sample size should be sufficient to support direct estimates of the population at the design group level. Therefore, a direct method of estimation is required that is both robust and efficient. A small bias would be acceptable if the overall error, accounting for the bias, is smaller than for unbiased options. The following section outlines the various options that were considered.

### *Options*

Four approaches were considered. The first uses simple Horvitz-Thompson estimation with the set of postcode counts adjusted by dual system estimation. This is in general unbiased with respect to repeated sampling but often not efficient. The second approach assumes the 2001 Census and the adjusted count for each postcode are proportional to each other. The sample data is used to estimate the constant ratio between the counts which is then used to predict the true count from the 2001 Census count for the non-sample postcodes. The third approach extends this to a simple linear relationship between the two counts of each postcode by the inclusion of a constant intercept. In addition to these three approaches a final option is considered which adjusts the ratio model to make it more robust to model failure. The modifications deal mainly with the problems of zero postcode counts and predicting for large postcodes.

### *Results*

The results clearly show that, of the three standard approaches, the ratio model is the best with a perfect CCS and when combined with Dual System Estimation. However, there are problems with the simple ratio model. It has a positive bias and is prone to making large over-estimates of the population. Simulations show that the robust ratio model is preferable as it protects against producing such estimates. The price is a slight negative bias, although total error is reduced. **It is this modified ratio model, combined with the Dual System Estimation methodology described previously, that is the preferred option for the production of design group estimates.**

# Estimation Strategy for Design Group Estimates by Age and Sex from the Census Coverage Survey

James Brown, Ray Chambers and Marie Cruddas

## 1) Introduction

There have been several papers (Brown *et al*, 1999<sup>1</sup> and Brown *et al*, 1999<sup>2</sup> for example) that have looked at the estimation strategy at the design group level. Brown *et al* (1999)<sup>1</sup> used regression estimation combined with dual system estimation to produce a set of age-sex estimates. The paper also examined the robustness of the approach to dependence between the census and the Census Coverage Survey. Brown *et al* (1999)<sup>2</sup> considered the ratio model combined with dual system estimation and found this to be a more satisfactory approach. It also looked at the effect of simple forms of matching error.

The production of high quality estimates by age and sex at the design group level, the design groups for England and Wales are defined in ONS(ONC(SC))00/10, is key to the whole 'One Number Census' process. This paper brings together and extends the earlier work to present an estimation strategy. The strategy combines dual system estimation with a modified ratio estimator. Simulation results show this to be a robust and efficient approach. Section 2 of this paper sets out the basic theory behind the different approaches that have been considered and Section 3 presents simulation results for those approaches. Section 4 then considers modifications to the ratio model to make it more robust while Section 5 contains a discussion of the paper.

## 2) Population Estimation Using the 2001 Census Coverage Survey

This paper assumes that the Census Coverage Survey (CCS) has been designed as outlined in Brown *et al* (1999)<sup>1</sup>. The important point is the fact that the design produces a sample of postcodes that are clustered within a random sample of enumeration districts (EDs) with five postcodes per ED, referred to as a cluster of postcodes. The choice of five postcodes per ED is a trade-off between cost efficiency and statistical efficiency based on simulations presented in ONS(ONC(SC))98/12. The enumeration districts are themselves a sample from the design group stratified by a ‘Hard to Count’ (HtC) index. The proposed formulation of the HtC index is outlined in ONS(ONC(SC))00/01 although this paper utilises an earlier prototype index. The aim of the index is to split EDs into strata based on their characteristics such that the resulting strata are more homogeneous with respect to the expected level of underenumeration.

The paper also assumes that both the census and CCS have been carried out and that the two databases have been successfully matched using the strategy also outlined in ONS(ONC(SC))00/01. In other words, for individuals counted in postcodes from the CCS sample it is possible to determine whether they were counted in both the census and the CCS, the census only, or the CCS only. There will also be some individuals who are not identified as they have been missed by both. Therefore the estimation strategy can be thought of in two parts. The first is to estimate the true population in the CCS postcodes and then use that to estimate for the whole design group. Section 2.1 considers the first part and Section 2.2 considers different approaches to the second part.

### 2.1) Dual System Estimation

A standard method for estimating underenumeration is Dual System Estimation. This was the approach used by the US Census Bureau following both the 1980 and 1990 US Censuses. Shortly after the census a Post-Enumeration Survey (PES), the US Census Bureau’s equivalent of the CCS, is used to obtain an independent re-count of the population in a sample of areas. Dual system estimation combines these two counts to estimate the true population, allowing for people missed by both the census and the PES, in the PES sample areas in the following way.

		CCS		
		Counted	Missed	
Census	Counted	$n_{11}$	$n_{10}$	$n_{1+}$
	Missed	$n_{01}$	$n_{00}$	$n_{0+}$
		$n_{+1}$	$n_{+0}$	$n_{++}$

(1)

$$\hat{n}_{++} = \frac{n_{1+} \times n_{+1}}{n_{11}}$$

Although the method is theoretically straightforward, in practice it has some problems.

- a) The DSE assumes that in the target population the matched PES and census counts follow a multinomial distribution. That is, the probabilities of being counted by either or both the PES and the census are **homogeneous** across the target population. This is unlikely for most populations.

- b) Unbiased estimation requires statistical **independence** between the census count and the PES count. This is impossible to guarantee.
- c) It is necessary to **match** the two data sources to determine whether individuals on the lists were counted once or twice. Errors in matching become biases in the dual system estimator (DSE).

In the 1990 Census the US Census Bureau tackled problem a) by splitting the population up into post strata based on factors (e.g. race) which were thought to affect an individual's probability of being counted, a method originally proposed by Sekar and Deming (1949). Problem b) is typically handled by operational procedures that ensure the operational independence of the census and the PES. Problem c) is essentially unavoidable but it is absolutely essential to ensure that errors due to matching are minimised. The work carried out for the 1990 US Census on all three problems is outlined in Hogan (1992, 1993).

If the US equivalent of the CCS covered the whole population the DSE defined by (1), under the assumptions already stated, would give an estimate of the total population in a particular post-strata. However, it only covers a sample so the US Census Bureau use a weighted DSE to estimate for the total population

$$\hat{N}_{++} = \frac{\sum_{i \in \text{PES}} n_{+1i} / \pi_i}{\sum_{i \in \text{PES}} n_{11i} / \pi_i} \times N_{1+} \quad (2)$$

where  $N_{1+}$  is the total census count for a particular post-strata and  $n_{+1i}$  is the PES count,  $n_{11i}$  is the matched individuals and  $\pi_i$  is the probability of inclusion for sampled area  $i$  within the post-strata. Further modifications are also applied to account for estimated overenumeration in the census due to erroneous enumeration and imputation at the processing stage.

This paper proposes a slightly different approach. Dual system estimation is thought of as a way of adjusting the sample counts generated by the CCS to account for those missed by the CCS. These adjusted counts are then used to estimate the total population, part two of the strategy. Assumption a) will be approximated by splitting the population into groups by age and sex within the sampled postcodes of each HtC category and therefore, the DSE will be done at a very low level of aggregation. This will, along with operational independence, also help ensure that assumption b) is well approximated. In addition, work presented in Brown et al (1999)<sup>1</sup> shows that the overall approach is robust to some dependence between the census and CCS and its robustness increases as the response rate achieved by the CCS increases.

As the proposal is to carry-out dual system estimation at a low level of aggregation the standard estimator (1) is corrected for its small sample bias to give

$$\hat{n}_{++} = \frac{(n_{1+} + 1) \times (n_{+1} + 1)}{(n_{11} + 1)} - 1 \quad (3)$$

This is the adjustment proposed by Chapman (1951) which is recommended by Seber (1982).

## 2.2) Models for Population Estimation Using the CCS

After the CCS there will be two population counts for each postcode in the CCS sample. One approach would be to assume that the CCS count is equal to the population count in the sampled postcodes and that, therefore, there is no underenumeration in the CCS. However, it is more sensible to assume that there will be underenumeration in both census and CCS and hence for each sampled postcode there are two counts, both with non-response. Under the assumptions of independence between the two counts and homogeneity of the census / CCS ‘capture’ probabilities for each age-sex group at the postcode level the DSE can be used to estimate the true population counts,  $Y_{aied}$ , for age-sex group  $a$  in postcode  $i$  from cluster  $e$  in HtC stratum  $d$ . The problem is then how to estimate the overall population total in the design area,  $T_a$ , for age-sex group  $a$  using this information.

### 2.2.1) Simple Approach.

The simplest approach to make such population estimates is to assume that information is only available for the census and the CCS in the sample areas. In this situation Alho (1994) proposes an adaptation of the Horvitz-Thompson estimator for  $T_a$  such that

$$\hat{T}_a = \sum_{i \in \text{CCS}} \frac{\hat{Y}_{aied}}{\pi_{ied}} \quad (4)$$

where  $\pi_{ied}$  is the probability of inclusion for postcode  $i$  from cluster  $e$  of HtC category  $d$  and  $\hat{Y}_{aied}$  is the corresponding DSE estimate for age-sex group  $a$ . An alternative is to compute the DSE at the cluster level giving

$$\hat{T}_a = \sum_{e \in \text{CCS}} \frac{\hat{Y}_{aed}}{\pi_{ed}} \quad (5)$$

where  $\pi_{ed}$  is the probability of inclusion for the cluster of postcodes  $e$  from HtC category  $d$  and  $\hat{Y}_{aed}$  is the corresponding DSE estimate for age-sex group  $a$ . As the postcode sample from within the sampled EDs / clusters is a simple random sample  $\pi_{ied}$  and  $\pi_{ed}$  are equal and therefore when the  $Y_{aied}$ 's (and by addition the  $Y_{aed}$ 's) are known without error both approaches give the same estimate for  $T_a$ . When dual system estimation is needed to estimate the  $Y$ 's, one would expect the second approach to be more stable due to larger counts being used to calculate the DSEs. However, it will possibly be slightly more susceptible to ‘correlation bias’ caused by the violation of assumptions a) and b).

### 2.2.2) Ratio Model for Population Estimation

In reality, census counts are available for all postcodes and can be used as auxiliary information to improve on the Horvitz-Thompson estimator. The simplest way to introduce these auxiliary data is to assume that the true count is proportional to the census count. For estimation this leads to the classical ratio model for each age-sex group. Dropping the age-sex group indicator  $a$ , and representing the census count in postcode  $i$  from cluster  $e$  of HtC stratum  $d$  by  $X_{ied}$ , this model can be written as

$$\begin{aligned} E\{Y_{ied} | X_{ied}\} &= R_d X_{ied} \\ \text{Var}\{Y_{ied} | X_{ied}\} &= \sigma_d^2 X_{ied} \\ \text{Cov}\{Y_{ied}, Y_{jfg} | X_{ied}, X_{jfg}\} &= 0 \text{ for all } i \neq j \end{aligned} \quad (6)$$

where  $R_d$  and  $\sigma_d^2$  are unknown parameters. Under the model in (6) it is straightforward to show (Royall, 1970) that the best linear unbiased predictor (BLUP) for the true population total  $T$  of an age-sex group is the stratified ratio estimator of this total given by

$$\hat{T}_{\text{RAT}} = \sum_{d=1}^5 \hat{R}_d \sum_{e=1}^{N_d} \sum_{i=1}^{M_e} X_{ied} \quad (7)$$

where  $N_d$  is the total number of EDs (clusters) in HtC stratum  $d$ ,  $M_e$  is the total number of postcodes in cluster  $e$ , and  $\hat{R}_d$  is an estimate of the population ratio of true to census counts. Strictly speaking the assumption in model (1) of zero covariance between postcodes counts is violated as the design of the CCS has the postcodes clustered. However, this is not a problem for estimation of the population total, as (2) remains unbiased when this assumption is violated with only a small loss of efficiency (Scott and Holt,

1982). Typically  $\hat{R}_d = \frac{\sum_{e=1}^{n_d} \sum_{i=1}^5 Y_{ied}}{\sum_{e=1}^{n_d} \sum_{i=1}^5 X_{ied}}$  where  $n_d$  is the number of clusters sampled in HtC index  $d$  and five is

the number of postcodes sampled from cluster  $e$ . In practice, of course, the  $Y_{ied}$  are unknown and replaced by their corresponding DSEs.

As with the Horvitz-Thompson approach the estimator of  $R_d$  can be adapted to allow for calculating the

DSE at the cluster level to give  $\hat{R}_d = \frac{\sum_{e=1}^{n_d} \hat{Y}_{ed}}{\sum_{e=1}^{n_d} X_{ed}}$ . Another additional alternative is to compute one DSE

across all the CCS sample postcodes within a HtC stratum and then ‘ratio’ this total up to a population estimate for that stratum by multiplying it by the ratio of the overall census count for the stratum to the census count for the CCS postcodes in the stratum. This final alternative is analogous to treating the HtC stratum as a post-stratum in the US Census context and applying the ratio estimator proposed by Alho (1994). However, if the true counts are known in the sampled areas (ie the CCS has no non-response) all three approaches using the ratio model are equivalent and give the same estimate for  $T_a$ .

One would expect the second and third approaches to have lower variances due to the larger counts contributing to the DSE but be increasingly subject to correlation bias due to heterogeneity of capture probabilities within each HtC stratum and possible dependence. Defining the HtC strata after the census can reduce this correlation bias as is done by the US Census Bureau. However, it appears unlikely that all the necessary data for such a post-stratification will be available in time for such an exercise to be carried out after the 2001 UK Census.

### 2.2.3) Regression Model for Population Estimation

The model in (1) forces a strictly proportional relationship between the census and true counts. Such a relationship is unlikely to be the case where census counts are close to zero, as will be the situation if estimation is carried out at the postcode level. Therefore, Brown *et al* (1999)<sup>1</sup> suggested the use of a

simple regression model to allow for the situation where the census counts for a particular postcode are close, but not equal to, zero. This model is given by

$$\begin{aligned}
 E\{Y_{ied} | X_{ied}\} &= \alpha_d + \beta_d X_{ied} \\
 \text{Var}\{Y_{ied} | X_{ied}\} &= \sigma_d^2 \\
 \text{Cov}\{Y_{ied}, Y_{jfg} | X_{ied}, X_{jfg}\} &= 0 \text{ for all } i \neq j
 \end{aligned}
 \tag{8}$$

Under (8) it is straightforward to show (Royall, 1970) that the best linear unbiased predictor (BLUP) for the true population total  $T$  of an age-sex group is then the stratified regression estimator

$$\hat{T}_{\text{REG}} = \sum_{d=1}^5 \sum_{e=1}^{N_d} \sum_{i=1}^{M_e} (\hat{\alpha}_d + \hat{\beta}_d X_{ied})
 \tag{9}$$

where  $\hat{\alpha}_d$  and  $\hat{\beta}_d$  are the OLS estimates of  $\alpha_d$  and  $\beta_d$  in (8). Like the ratio estimator (7), (9) is robust to the correlation of postcodes due to the sample design (Scott and Holt, 1982). Unfortunately, it is not robust to a large number of zero census / CCS counts, since the fitted regression line can then be significantly influenced by the large number of sample points at the origin.

### 3) Simulation Study

To assess the performance of the three estimators of the population total described in Section 2.2 when the CCS design described in Brown *et al* (1999)<sup>1</sup> is applied to a population a simulation study was undertaken and is described in this section. Anonymised individual records for a local administrative area from the 1991 Census augmented by a HtC index are used as the basis for the simulation. The population is treated as a design area and has approximately 450,000 individuals within 170,000 households. It has over 10,000 postcodes and there are 930 enumeration districts.

#### 3.1) Applying the CCS Design to the Simulation Population

As already stated it is expected that underenumeration in the 2001 Census will be higher in areas characterised by particular social, economic and demographic characteristics. For example, people in dwellings occupied by more than one household (multi-occupancy) have a relatively high probability of not being enumerated in a census. This heterogeneity is accounted for by stratifying the enumeration districts (and hence the postcodes contained within them) by a ‘Hard to Count’ (HtC) index. The “prototype” HtC index used here is based on a linear combination of the variables:

- percentage of heads of household who experienced language difficulty;
- percentage of young people who migrated into the enumeration district in the last year;
- percentage of imputed residents for the enumeration district;
- percentage of households in multiply-occupied buildings; and
- percentage of households which were privately rented.

The distribution of the enumeration districts in the simulation population by this HtC index is given in Table 1.

Table 1 also shows the number of EDs selected in each stratum. Selection was carried out using size stratified sampling based on the design described in Brown *et al* (1999)<sup>1</sup>, leading to a total of 35 EDs being selected. A simple random sample of five postcodes (or less if the selected ED does not contain five postcodes) is then selected from each sampled ED.

TABLE 1

*Distribution of enumeration districts by HtC index with first stage sample*

HtC Index Value	Number of Enumeration Districts	Sample of Enumeration Districts
Very Easy	144	6
Easy	210	7
Medium	186	6
Hard	193	7
Very Hard	197	9
<b>TOTAL</b>	<b>930</b>	<b>35</b>

### 3.2) Simulating a Census and its CCS

Census underenumeration was simulated by each individual in the population being given a probability of being counted in a census. These probabilities depend on individual characteristics and are based on research by the 'Estimating With Confidence Project' (Simpson *et al*, 1997) following the 1991 Census. In particular, there is considerable variation in the individual probabilities by age and sex. They also vary by HtC index; the census variable 'Primary Activity Last Week'; and there is also a small enumeration district effect. However, the probabilities are still heterogeneous even when all these factors are taken into account. Whole households are also assigned probabilities of being counted in the census. These are based on averaging the individual probabilities associated with the adults within the households. Household probabilities also vary according to the tenure of the household and the household size. The household and individual probabilities remain fixed throughout the simulation study.

Each individual and household is also assigned a factor that defines the differential nature of response in the CCS. These mirror the same pattern as the census probabilities but the differentials are less extreme. This extends the simulation study in Brown *et al* (1999)<sup>1</sup> so that there is some heterogeneity in both the census and the CCS for age-sex groups at the postcode level.

To generate a census and its corresponding CCS, independent Bernoulli trials are used to determine first whether the household is counted and second whether the individuals within a counted household are counted. There is also a check that converts a counted household to a missed household if all the adults in the household are missed. In these simulations the census and CCS outcome for households and individuals are independent. This assumption can be investigated by specifying the odds ratio between the two outcomes to be different from one, see Brown *et al* (1999)<sup>1</sup>. Two levels of coverage are used in the CCS. First a perfect CCS is simulated and then coverage is set at approximately 90 per cent for households with 98 per cent of individuals within those households being counted. For each census ten CCS postcode samples are selected based on the design in Table 1. The estimators described in Section 2.2 are then applied to each age-sex group and population totals are calculated. The whole process is repeated for 100 independent censuses.

### 3.3) Population Estimation Results

For the simulation of 100 censuses the average census coverage is 94.90 per cent which drops to 15 per cent for males aged 20-24. Details of the census coverage for each age-sex group are given in Appendix I. The overall coverage is rather less than 1991 where it was around 98 per cent and aims to assess the robustness of the procedure to increased probabilities of underenumeration. The six estimators being evaluated are:

- 1) The Horvitz-Thompson (HT) estimator with the DSE at the postcode level
- 2) The Horvitz-Thompson (HT) estimator with the DSE at the cluster level
- 3) The ratio estimator with the DSE at the postcode level
- 4) The ratio estimator with the DSE at the cluster level
- 5) The ratio estimator with the DSE at the HtC index level
- 6) The regression estimator with the DSE at the postcode level

As this is a simulation calculating the relative root mean square errors (RRMSE) and the relative biases can be used to assess the performance of the estimators relative to each other (and the census) over the 1000 CCSs. For each estimator the RRMSE is defined as:

$$\text{RRMSE} = \frac{1}{\text{truth}} \times \sqrt{\frac{1}{1000} \times \sum_{j=1}^{1000} (\text{observed}_j - \text{truth})^2} \times 100 \quad (10)$$

and can be considered as a measure of the total error due to bias and variance. Relative bias is defined as:

$$\text{Relative Bias} = \frac{1}{\text{truth}} \times \frac{1}{1000} \times \sum_{j=1}^{1000} (\text{observed}_j - \text{truth}) \times 100 \quad (11)$$

Bias in an estimator is usually considered a poor feature, as it cannot be estimated from the sample. However, it can be better overall to adopt a slightly biased estimator if its total error is small.

TABLE 2

*Performance of the population estimators for the population total*

Type of Estimator	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Horvitz-Thompson			
Perfect CCS	0.24	6.83	1.11
Postcode DSE	0.11	6.82	0.50
Cluster DSE	0.22	6.82	1.03
<b>Ratio Estimator</b>			
<b>Perfect CCS</b>	<b>0.24</b>	<b>0.52</b>	<b>15.92</b>
<b>Postcode DSE</b>	<b>0.10</b>	<b>0.49</b>	<b>6.70</b>
<b>Cluster DSE</b>	<b>0.22</b>	<b>0.53</b>	<b>13.95</b>
<b>Index DSE</b>	<b>0.23</b>	<b>0.54</b>	<b>15.30</b>
Regression Estimator			
Perfect CCS	0.37	0.64	22.77
Postcode CCS	0.23	0.57	13.97

Table 2 summarises the results for the estimation of the total population by summing the individual age-sex estimates. Across the different types of estimator the relative bias for each is similar although the bias estimated from the simulations for the Horvitz-Thompson estimators is not statistically significant. However, when you look at the total error measured by the relative RMSE the estimators using the Horvitz-Thompson approach are much less efficient. This is exactly what you would expect as the Horvitz-Thompson estimators make no use of extra information available from the 2001 Census for postcodes not in the CCS sample.

Considering the ratio and regression based estimators those using the ratio model are marginally better both in terms of bias and total error. The significant bias for the ratio model with a perfect CCS is due to model failure for a particular. On the face of it the postcode DSE with the ratio model looks ‘best’, as the relative bias of this estimator is less than for a perfect CCS. However, caution is needed as the reduction in bias when the DSE component is introduced at the postcode level suggests that dual system estimation introduces an additional negative bias. It is not the case that this fixes the problems causing the bias for a perfect CCS. The results in Table 2 suggest this additional bias is not introduced when the DSE is used at the cluster or index level.

TABLE 3  
*Performance of the DSE at two levels for estimating the sample population*

	Relative Bias (%)	Z-value for Bias
DSE at Postcode Level	-0.10	22.01
DSE at Cluster (5 Postcode) Level	0.0086	1.82

Table 3 presents results for just estimating the population in the sample postcodes (by simply summing the DSEs) over the simulation. It demonstrates that at the postcode level the DSE is not working correctly due to very small counts in the individual age-sex groups. Indeed, it does have a highly significant negative bias. However, the DSE at the cluster level is essentially unbiased over the simulation. Therefore, the cluster level DSE with the ratio model looks the ‘best’ option as a compromise between the unstable DSE at the postcode level due to very small (zero) counts and the increased risk of correlation bias with the index level DSE.

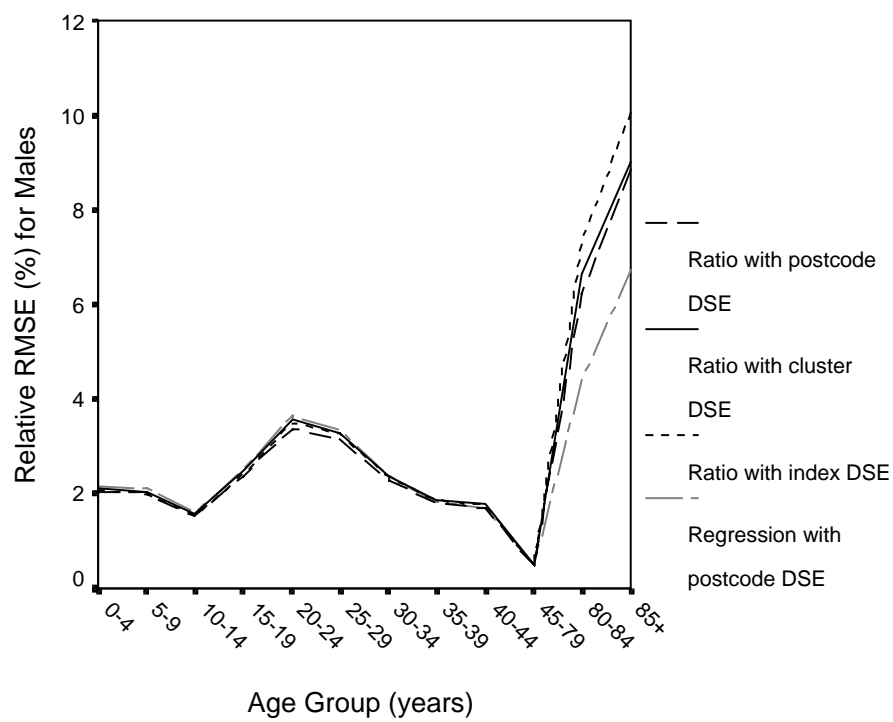
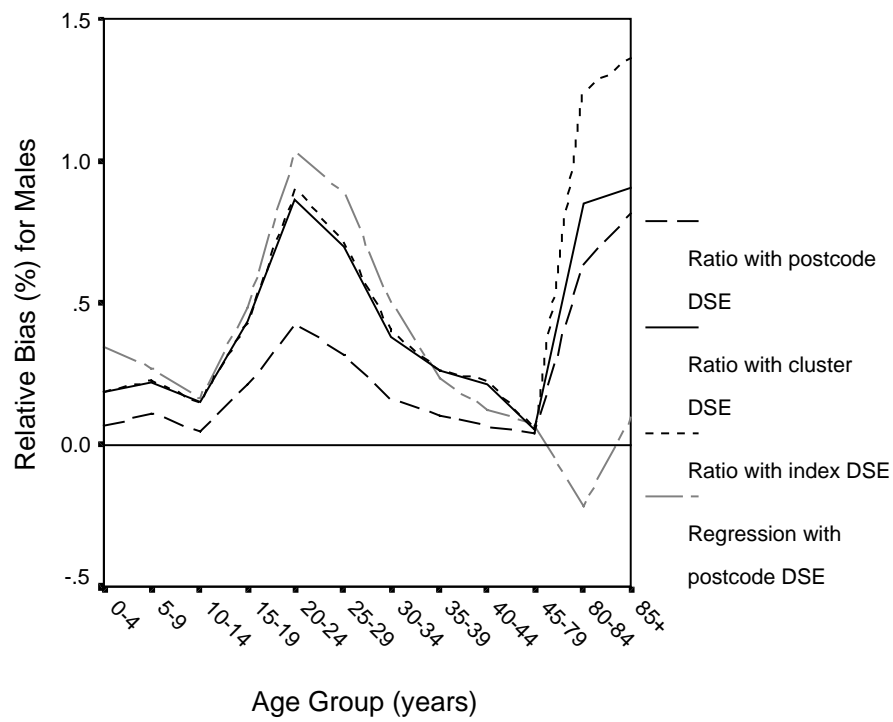
Looking at the total can hide problems with the estimation of the individual age-sex population totals. Figure 1 presents the results for the male age groups using the ratio and regression models. The Horvitz-Thompson estimator results are in Appendix II.

Figure 1 clearly shows that across the age groups for males the four estimators are very similar in terms of RRMSE with the exception of men aged 85 years and over where there are more noticeable differences. With respect to bias the regression estimator with postcode DSE has a higher bias for the young age groups. This is what gives the regression estimator its high relative bias in Table 2. The three estimators based on the ratio model have a higher relative bias for the oldest age groups. However, in terms of the population total these are small groups and so do not impact on the results in Table 2. Of the three estimators based on the ratio model the postcode DSE has a consistently lower bias. This reflects the negative bias in the DSE at this level demonstrated by Table 3. In general it is not good practice to rely on biases cancelling each other to get a ‘better’ estimator and in terms of total error there is very little impact suggesting that reductions in bias are balanced by increased variance.

The estimators based on the ratio model are better with little to choose between the cluster level DSE and the index level DSE. However, the estimator using the index level DSE needs to be treated with care as it relies heavily on the HtC index defining homogeneous strata. In the simulation this is the case once age

and sex are also controlled for. However, in 2001 this assumption will be shakier when the index has been defined for postcodes based on their 1991 characteristics and there will certainly be postcodes that will have changed in ten years. This will cause the DSE calculated at the HtC stratum level to be biased. For the postcode-based estimator this will not impact on the individual DSEs to cause bias but it will increase the variance as the relationship between the census and CCS counts within each stratum will not be as strong. In addition, at this level of aggregation the DSE is unstable. The cluster level estimators are, as already stated, a good compromise between the two 'extremes'.

Figure 1: Ratio and regression estimators combined with dual system estimation for males by age



## 4) A ‘Robust’ Estimation Strategy

The results from the simulation presented in Section 3.3 demonstrate the existence of problems with both the ratio and the regression model as the census count gets small causing model failure and potentially bias. As stated in Section 2.2 the regression model will fit well when census counts are approaching zero and the CCS is finding extra people but it will not be robust to a large number of origin points. As the postcode is a very small geographic area the count for a particular age-sex group will often be zero. While origin points do not affect the ratio model, as it is constrained to pass through the origin, postcodes where the census count is zero and the CCS is greater than zero do. These happen in a few postcodes for all the age-sex groups.

There is a second issue that impacts on the estimation. The ratio estimator essentially uses the ratio estimated from the sample to predict the count in the non-sample areas. This becomes a problem when there are census counts in the non-sample postcodes that are greater than those in the sample postcodes. In that situation the danger is making a prediction where there is no sample to support such a prediction and a few outlying census counts can have a considerable impact on the final estimate.

There is a third situation that can occasionally occur. It happens when there are large numbers of census counts in the sample areas that are zero, some with a non-zero CCS count, combining with an extreme form of problem two where the non-zero census counts in the sample areas are all close to zero. This results in the situation where zero census non-zero CCS counts have a large impact on the estimated ratio which is then used to predict for counts well outside the range of the sample data. The oldest age groups are particularly vulnerable to this happening leading to the positive bias in Figure 1.

### 4.1) Models for Robust Estimation

The previous section highlights three problems with the ratio model that are causing model mis-specification bias. This section takes each problem in term and proposes a strategy to deal with it. The first problem is a zero census count with a non-zero CCS count. Brown *et al* (1999)<sup>2</sup> proposed a mixture type model to cope with this problem but simulations showed it to be difficult to estimate all the necessary parameters. For this paper the following simpler approach is proposed.

$$\begin{aligned}
 &\text{If } X_{ied} > 0; \\
 &E\{Y_{ied} | X_{ied}\} = R_d X_{ied} \\
 &\text{Var}\{Y_{ied} | X_{ied}\} = \sigma_d^2 X_{ied} \\
 &\text{Cov}\{Y_{ied}, Y_{jfg} | X_{ied}, X_{jfg}\} = 0 \text{ for all } i \neq j
 \end{aligned} \tag{12}$$

$$\begin{aligned}
 &\text{If } X_{ied} = 0; \\
 &E\{Y_{ied}\} = \mu_d \\
 &\text{Var}\{Y_{ied}\} = \sigma_d^2 \\
 &\text{Cov}\{Y_{ied}, Y_{jfg}\} = 0 \text{ for all } i \neq j
 \end{aligned} \tag{13}$$

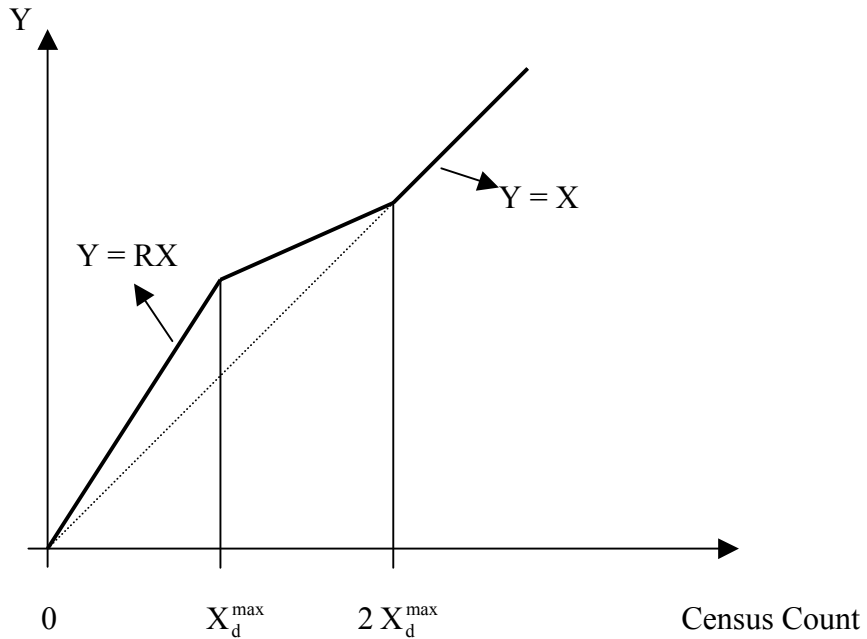
It works by splitting the estimation into two parts, one for postcodes with a zero census count (13) and one for postcodes with a non-zero census count (12). Here, the model proposed for the zero census counts is just the simple homogeneous model and for the non-zero counts the standard ratio model.

The second problem is the prediction outside the range of the sample data. Empirical evidence from the simulations suggests that this is important and causes large ‘over-estimates’. To reduce the bias caused by this the ratio part of the estimator generated by (12) and (13) is modified such that the overall estimator is given by

$$\hat{T}_d = \sum_{X_{ied}=0} \hat{\mu}_d + \sum_{0 < X_{ied} \leq x_d^{\max}} \hat{R}_d X_{ied} + \sum_{x_d^{\max} < X_{ied} \leq 2x_d^{\max}} \{(2 - \hat{R}_d)X_{ied} + 2x_d^{\max}(\hat{R}_d - 1)\} + \sum_{X_{ied} > 2x_d^{\max}} X_{ied} \quad (14)$$

$$\hat{T} = \sum_{d=1}^5 \hat{T}_d$$

where  $x_d^{\max}$  is the largest census count for a CCS sample postcode in a particular age-sex HtC combination and  $\hat{\mu}_d$  is estimated by the simple mean of the sample data with  $X_{ied} = 0$ . Graphically, these modifications to the ratio part of the estimator (14) can be represented as



where the aim is to reduce the influence of outliers in the census on the final estimate of  $\hat{T}$ .

The final problem does not occur often but when it does the result is usually a dramatic over-estimate. The older ages are most vulnerable but it can occur for any age-sex group within a HtC category for a given sample. To combat this, in the situations where there are not three distinct non-zero census counts in the CCS sample the estimation strategy outlined by (12)-(14) is not used for that particular age-sex HtC combination. Instead, an alternative model is used given by

$$\begin{aligned}
E\{Y_{ied} | X_{ied}\} &= X_{ied} + \mu_d \\
\text{Var}\{Y_{ied} | X_{ied}\} &= \sigma_d^2 \\
\text{Cov}\{Y_{ied}, Y_{jfg} | X_{ied}, X_{g_{jfg}}\} &= 0 \text{ for all } i \neq j
\end{aligned} \tag{15}$$

which is just a regression model where  $\Xi_d$  is constrained to one. The resulting estimator of the total for the particular age-sex HtC combination is then given by

$$\hat{T}_d = \sum_{e=1}^{N_d} \sum_{i=1}^{M_e} (X_{ied} + \hat{\mu}_d) \tag{16}$$

where  $N_d$  and  $M_e$  are defined as before and  $\hat{\mu}_d = \bar{y}_d - \bar{x}_d$  where  $\bar{y}_d$  and  $\bar{x}_d$  are the unweighted sample means. The justification of (15) in this situation is that the model does not attempt to estimate a slope parameter from very little information. However, it does utilise the fact that the CCS has identified some extra people and combines this with the fact that census counts are available for all postcodes.

## 4.2) Applying the Strategy

The same simulation as used in Section 3 can now be applied to the robust estimation strategy. For a perfect CCS the above strategy can be applied directly to the sample of postcode counts generated by the simulation. However, in reality  $Y_{ied}$  will not be known but will be estimated using dual system estimation. It has already been shown that at the postcode level the DSE is unsatisfactory but for prediction purposes the models and estimators proposed in Section 4.1 are at the postcode and not cluster level. Therefore, the postcode level DSE is used but scaled to the cluster level DSE so that they do sum to an unbiased estimate of the population in the sample postcodes. In other words, if  $Y_{ied}$  is the raw CCS count and  $B_{ied}$  is the matched count for postcode  $i$  from cluster  $e$  of HtC category  $d$  then

$$Y_{ied}^{\text{DSE}} = \frac{Y_{ied} \times X_{ied}}{B_{ied}} \quad Y_{ed}^{\text{DSE}} = \frac{\sum_{i=1}^5 Y_{ied} \times \sum_{i=1}^5 X_{ied}}{\sum_{i=1}^5 B_{ied}} \quad \rho_{ed} = \frac{Y_{ed}^{\text{DSE}}}{Y_{ied}^{\text{DSE}}} \quad \hat{Y}_{ied} = \rho_{ed} Y_{ied}^{\text{DSE}} \tag{17}$$

where  $Y_{ied}^{\text{DSE}}$  is the postcode level DSE,  $Y_{ed}^{\text{DSE}}$  is the cluster level DSE, and  $\rho_{ed}$  scales  $Y_{ied}^{\text{DSE}}$  so that  $\hat{Y}_{ied}$ , the count used for estimation, is consistent with  $Y_{ed}^{\text{DSE}}$ .

In addition to the strategy outlined in Section 4.1 the program is also coded to remove postcodes as outliers if the ratio defined for that postcode exceeds pre-specified bounds when using the model given in (12) in conjunction with the estimator defined by (14). The bounds are; greater than or equal to three for HtC categories one and two, greater than or equal to four for HtC categories three and four, and greater than or equal to five for HtC category five. Currently, these bounds have been chosen based on examining output generated by the simulation. There is the possibility of further

work to refine the bounds and make them age-sex and HtC specific. If a postcode is defined as an outlier it is removed from the estimation process and then simply added on to the estimate of T at the end.

### 4.3) Results from the Robust Estimation Strategy

The simulation was repeated using the methodology described in Section 3 with the same census and CCS coverage rates. In this case there are just two estimators being evaluated:

- 1) The ‘robust’ ratio estimator with the DSE at the postcode level
- 2) The ‘robust’ ratio estimator with the DSE at the postcode level constrained to the DSE at the cluster level

where ‘robust’ ratio estimator refers to the whole strategy outlined in Section 4.1. Table 4 summarises the results for the estimation of the total population by summing the individual age-sex estimates.

TABLE 4

*Performance of the ‘robust’ ratio estimators for the population total*

Type of Estimator	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Ratio Estimator			
<b><i>Perfect CCS</i></b>	<b><i>-0.02</i></b>	<b><i>0.47</i></b>	<b><i>-1.30</i></b>
Postcode DSE	-0.32	0.56	-21.71
<b><i>Postcode DSE Constrained to Cluster DSE</i></b>	<b><i>-0.17</i></b>	<b><i>0.50</i></b>	<b><i>-11.15</i></b>

Table 4 demonstrates that for a perfect CCS the adjustments to the estimation strategy are working as expected to reduce bias. This can be seen by comparing  $-0.02$  in Table 4 with  $0.24$  in Table 2. In addition, the bias in Table 4 is no longer significantly different from zero. Once the DSE is introduced, as expected the results using the unadjusted postcode level DSE are not so good and the corrections to the ratio model along with negative bias and instability in the DSE estimates ‘over-corrects’ from  $0.10$  in Table 2 to  $-0.32$  in Table 4. However, constraining the postcode DSEs as described in Section 4.1 has helped to combat this. The  $-0.17$  in Table 4 can be compared to the cluster level DSE result in Table 2 of  $0.22$ . There is still an ‘over-correction’ and the z-value suggests that it is significant but there is also a decrease in the total error from  $0.53$  in Table 2 to  $0.50$  in Table 4. This not only represents the reduction in absolute bias but also a reduction in variance.

As before, looking at the total can hide what is happening across the age-sex groups. Figure 2 presents results for males that compares the robust approach using the postcode level DSE constrained to the cluster level DSE with the standard ratio model using cluster level DSE. The graph for relative RMSE shows a gain at all ages in terms of total error from using the robust approach. The graph for the bias shows that the robust procedures do introduce some negative bias into the estimator, particularly at the youngest age groups. However, it also shows that the procedures are doing well

to correct the large positive bias for males in the age groups 20-34 that is present with the standard estimators. The robust procedures also work better for the oldest age groups in terms of bias and total error where the standard ratio model is particularly unsatisfactory due to very low population counts in many postcodes.

Figure 3 demonstrates the additional advantage of the robust procedures. The reduction in total error reflects not only a reduction in absolute bias but also a reduction in variance. Figure 3 shows that this is achieved because the robust strategy stops the estimator producing the large over-estimates of the population by reducing the influence of outlying points in the estimation procedures.

Figure 2: Comparison of the robust ratio model using a postcode DSE constrained to a cluster DSE with the standard ratio model using cluster DSE for males by age

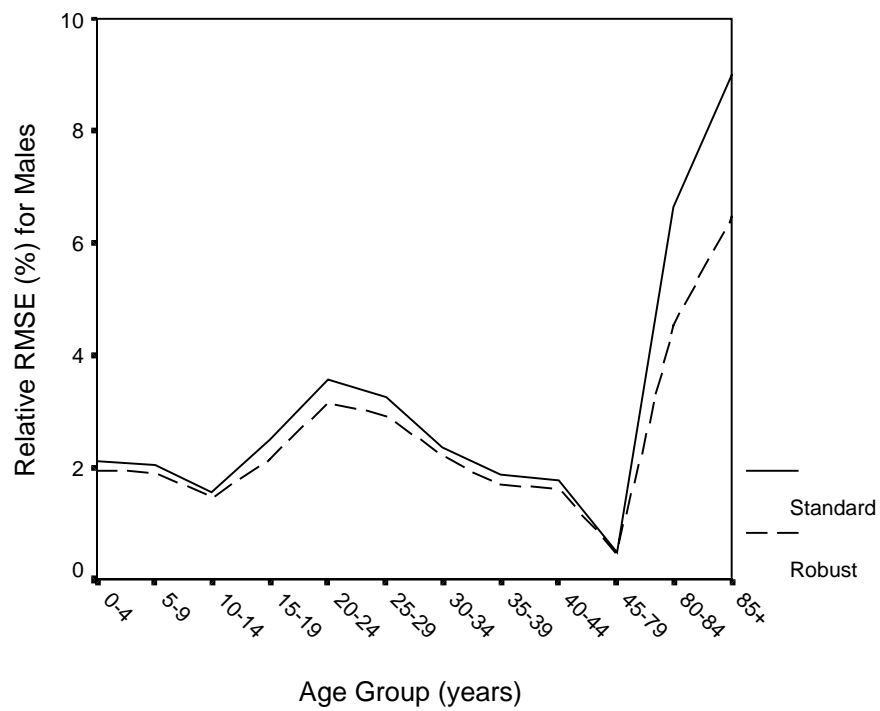
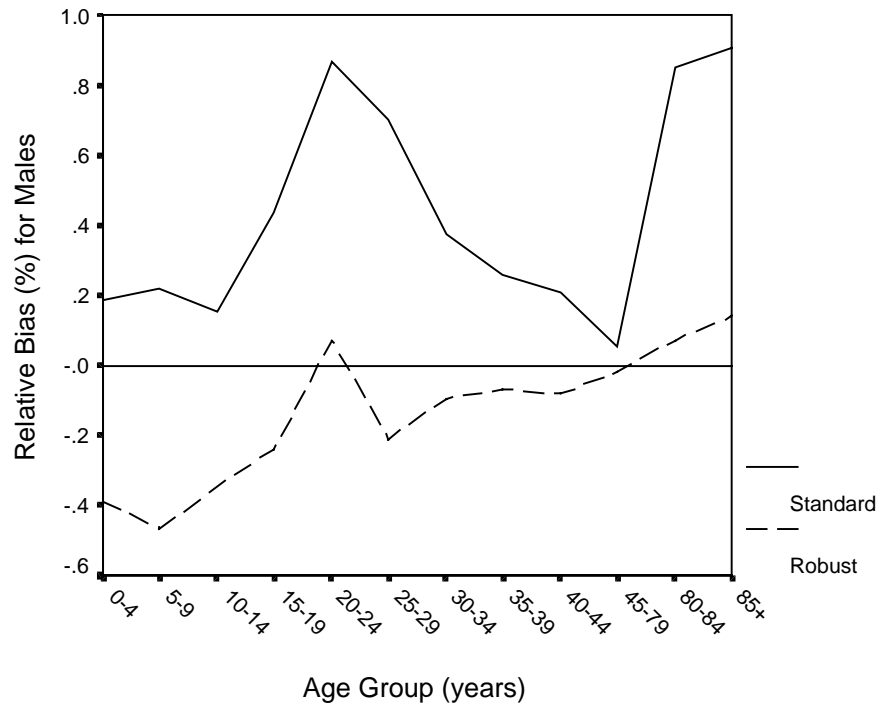
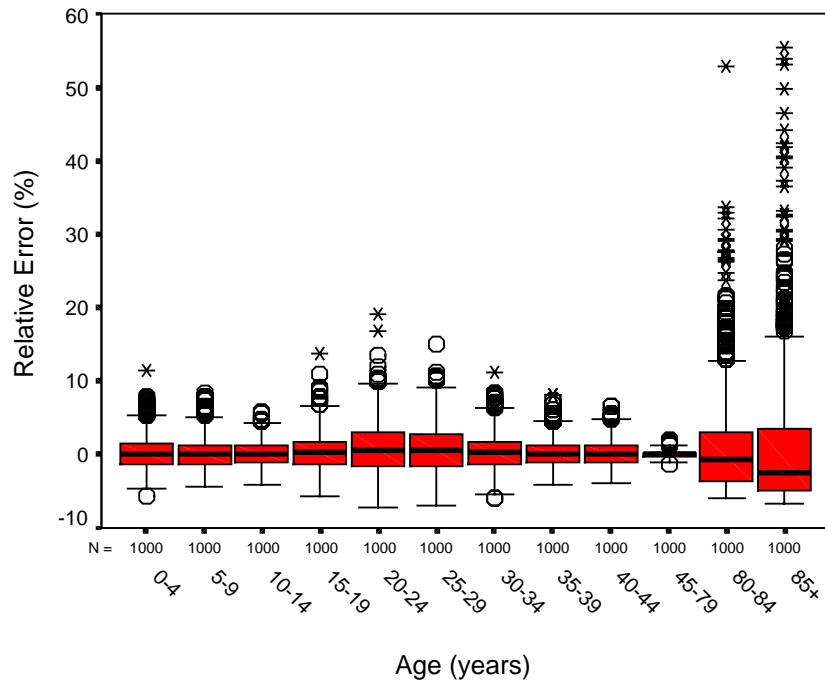
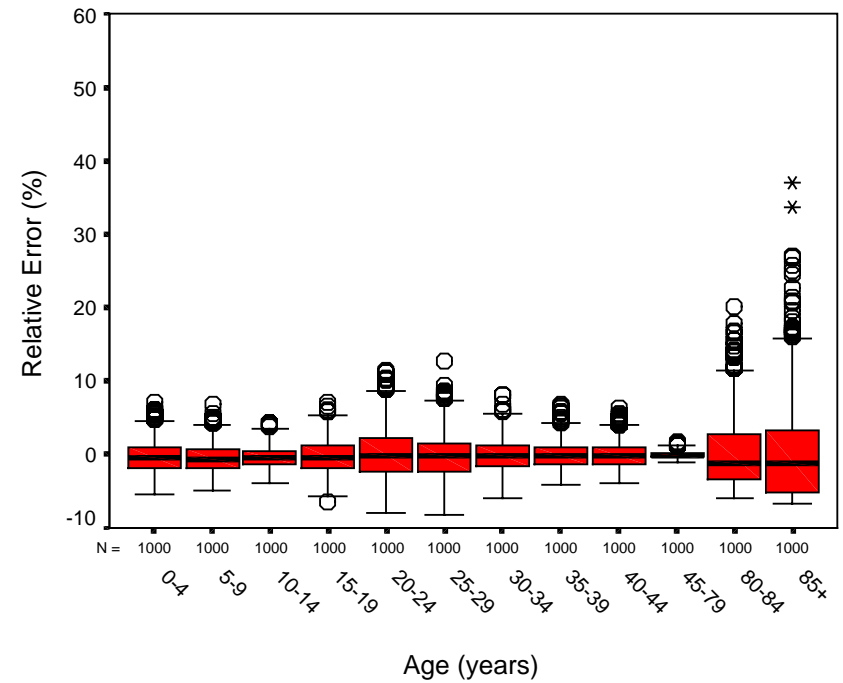


Figure 3: Distributions of the errors for the standard and robust strategies

*Errors from a standard ratio estimator with a cluster level DSE for males by age*



*Errors from a robust ratio estimator with a constrained postcode level DSE for males by age*



## 5) Discussion

The work presented in this paper has shown that the estimation strategy of combining dual system estimation with ratio / regression estimation techniques works well. Some problems have been highlighted. The DSE has a negative bias when produced for very small areas of aggregation, such as the postcode by age-sex group. The standard ratio and regression estimators suffer from a positive bias due to model failure and outliers. The proposed robust strategy constrains the postcode DSE to the cluster level DSE to ensure that there is little or no bias in the estimator from the DSE. The adjustments to correct for the model bias of the standard ratio estimator tend to give a slight negative bias. However, this is considered a price worth paying to protect against the strategy giving large over-estimates due to outliers in the sample.

The results presented here are based on a simulation study and in particular they have used a prototype HtC index that has five levels. Current methodology for the HtC index presented in ONS(ONC(SC))00/01 has a three level index that is calculated slightly differently to the prototype used here. The impact of this is to increase sample sizes in each HtC stratum, which will improve the estimation of the relationship between the census and CCS counts.

## References

- Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics* **10**, 245 - 256.
- 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.
- Brown, J. J., Diamond, I. D., Chambers, R. L., and Buckner, L. J. (1999)<sup>2</sup> The role of dual system estimation in the 2001 Census coverage surveys of the UK. *Population Association of America Annual Conference*, 23<sup>rd</sup> to 25<sup>th</sup> March 1999, New York.
- Chapman, D. G. (1951) Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Public. Stat.* **1**, 131-160.
- Hogan, H. (1992) The 1990 post-enumeration survey: an overview. *The American Statistician* **46**, 261-269.
- Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association* **88**, 1047-1060.
- ONS(ONC(SC))98/12 (1998) Census Coverage Survey: Precision of population estimates for different sample sizes and design areas.
- ONS(ONC(SC))00/01 (2000) One Number Census Methodology.
- ONS(ONC(SC))00/03B (2000) One Number Census Local Authority Estimation.
- ONS(ONC(SC))00/10 (2000) Design groups for 2001.
- Royall, R. M. (1970) On finite population sampling under certain linear regression models. *Biometrika* **57**, 377-387.
- Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848-854.
- Seber, G. A. F. (1982) The estimation of animal abundance and related parameters. Second edition published by *Charles Griffin & Company Ltd*, London.
- Sekar, C. C. and Deming W. E. (1949) On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44**, 101-115.
- Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends* **90**, 31-39.

Appendix I – Census coverage by age-sex group for the 100 simulated censuses

Sex	Age (years)	Mean Census Coverage (%)
Male	0-4	92.74
Male	5-9	93.79
Male	10-14	95.65
Male	15-19	91.23
Male	20-24	84.09
Male	25-29	85.47
Male	30-34	91.31
Male	35-39	95.12
Male	40-44	95.95
Male	45-79	98.38
Male	80-84	94.87
Male	85+	94.91
Female	0-4	93.37
Female	5-9	94.65
Female	10-14	96.42
Female	15-19	94.90
Female	20-24	92.23
Female	25-29	93.12
Female	30-34	96.28
Female	35-39	98.02
Female	40-44	97.72
Female	45-79	98.62
Female	80-84	91.62
Female	85+	84.26

Appendix II – Results for the Horvitz-Thompson estimator for males by age

