



ONS(ONC(SC))99/07

ONE NUMBER CENSUS STEERING COMMITTEE

The Role of Dual System Estimation in the 2001 Census Coverage Surveys of the UK

1. Attached is a copy of the paper that was presented at the Population Association of America Conference held in New York, 25 – 27 March 1999. This paper is circulated for information.
2. **The Steering Committee are asked to note the paper.**

**Lisa Buckner
Census Division
Office for National Statistics
Room 4200W
Segensworth Road
Titchfield
Fareham
HANTS
PO15 5RR**

June 1999

The Role of Dual System Estimation
in the
2001 Census Coverage Surveys of the UK

James Brown, Ian Diamond, and Ray Chambers
University of Southampton, UK

Lisa Buckner
Office for National Statistics, UK

Paper presented at the Population Association of America Annual Conference,
New York, 27th March 1999. Not to be referenced or reproduced without the
prior permission of the authors.

Address for Correspondence:

James Brown
Department of Social Statistics
University of Southampton
Southampton SO17 1BJ, UK
jjb1@soton.ac.uk

1) Introduction

Most countries conduct population censuses. By definition a census is a ‘complete’ count of some well-defined population at a point in time. Governments use their population censuses (usually every ten years) as the basis for planning and resource allocation down to very small areas. Accurate population counts for these areas are therefore extremely important. However, an increasing problem for many countries is what to do when the census, the gold standard, suffers from underenumeration. The United Kingdom (UK) has faced this problem for a number of censuses with net underenumeration measured by a post enumeration survey (PES). In 1991, though, the net underenumeration suggested by the PES was rather less than that suggested, at the national level, by demographic estimates. As a result it was not possible to allocate underenumeration below the national level using the PES and a demographic strategy was developed (Diamond, 1993 and Simpson *et al*, 1997). To ensure that this does not occur in 2001 a major research project has been undertaken so that, in 2001, the Office for National Statistics (ONS) will be in a position to **estimate** and **adjust** accurately census outputs for net underenumeration. The ultimate aim is to adjust the actual census database and create a ‘One Number Census’ (ONC) so that all tabulations reflect the estimated underenumeration and all figures reported by the census are consistent. However, before this can be done, estimates of the population at sub-national administrative areas, the level at which most resource allocation takes place, are required. It is this estimation problem which is the focus of this paper. The overall ONC research is described in Brown *et al* (1999).

1.1) 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 PES 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. 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). However, Alho *et al* (1993) show that the post-stratification employed in the 1990 US Census did not completely solve a) and they propose a logistic model to avoid this. 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).

Generalisation of the DSE counts from the sampled PES areas to the whole population can be carried out using a variety of survey estimation methods. In this paper DSE methodology is combined with ratio and regression estimation to achieve this aim. A series of estimators are proposed and evaluated using a simulation study. The robustness of the regression approach is examined with respect to a simple form of problem c) and problem a). Work already carried out (Brown *et al*, 1999) suggests that in the absence of a) and c) this approach behaves as one would expect in the presence of dependence. However, it is robust to the census and PES not being completely independent and its robustness increases as the response rate achieved in the PES increases.

2) Population Estimation using the 2001 Census Coverage Survey

2.1) Design of the Census Coverage Survey (CCS)

In the UK in 2001 the PES will concentrate only on coverage (as opposed to both quality and coverage) using a short questionnaire and large sample size. This will be known as the Census Coverage Survey (CCS). The aim of the CCS is to estimate population totals by 24 age and sex categories for groups of local administrative areas with total expected populations of around 0.5 million. For each sex there are five year age groups up to 40 to 44 year olds with 45 to 79 year olds as one group, 80 to 84 year olds, and those over 85 years of age. The 45 to 79 year olds have been combined in the research as there was little evidence of any underenumeration for these people in 1991. However, the methodology does not depend on this grouping and it will be reviewed in 2001. The design is based on a set of auxiliary variables, the 1991 census counts, for each age-sex group by enumeration district within a design group.

The CCS will be carried out between four and six weeks after the 2001 Census. It will be based on a two-stage sample design. At the first stage a stratified random sample of enumeration districts (EDs) will be drawn within a design area. An ED is an area containing about 200 households and represents the workload for one Census enumerator. Each design area is a collection of (usually contiguous) local administrative areas with an average 1991 population of around 500,000 persons. Within a design area the enumeration districts are stratified by a hard to count (HtC) index with categories $d = 1$ to D . The index is formed using 1991 data. It is based on variables that represent the social, economic, and demographic characteristics associated with people who were considered hard to count in the 1991 census. At the design stage its role is to ensure that all types of EDs are sampled. Further size stratification of EDs, within each level of the HtC index, based on the 1991 census counts improves efficiency by reducing within stratum variance.

The second stage of the CCS design will consist of the random selection of a fixed number of small areas called postcodes (containing on average fifteen households) within each selected enumeration district. All households in the selected CCS postcodes will then be independently enumerated using a short personal interview which will ascertain the household structure at the time of the census.

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_{aid} , for age-sex group a in postcode i 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) Ratio Model for Population Estimation

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 . However, census counts are available for all postcodes and can be used as auxiliary information to improve this 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 of HtC stratum d by X_{id} , this model can be written as

$$\begin{aligned} E\{Y_{id} | X_{id}\} &= R_d X_{id} \\ \text{Var}\{Y_{id} | X_{id}\} &= \sigma_d^2 X_{id} \\ \text{Cov}\{Y_{id}, Y_{jf} | X_{id}, X_{jf}\} &= 0 \text{ for all } i \neq j \end{aligned} \quad (1)$$

where R_d and σ_d^2 are unknown parameters. Under the model in (1) 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

$$\hat{T}_{\text{RAT}} = \sum_{d=1}^5 \hat{R}_d \sum_{i=1}^{N_d} X_{id} \quad (2)$$

where N_d is the total number of postcodes in HtC stratum d 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 $\hat{R}_d = \frac{\sum_{S_d} Y_{id}}{\sum_{S_d} X_{id}}$ where S_d represents the sampled postcodes

in HtC index d . In practice, of course, the Y_{id} are unknown and replaced by their corresponding DSEs.

An alternative to this estimator 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 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). One would expect this second approach to have a lower variance due to the larger counts contributing to the DSE but be subject to more bias due to heterogeneity of capture probabilities within each HtC stratum. Defining the HtC strata after the census can reduce this bias. 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.

2.2.2) 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) 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

$$\begin{aligned} E\{Y_{id} | X_{id}\} &= \alpha_d + \beta_d X_{id} \\ \text{Var}\{Y_{id} | X_{id}\} &= \sigma_d^2 \\ \text{Cov}\{Y_{id}, Y_{jf} | X_{id}, X_{jf}\} &= 0 \text{ for all } i \neq j \end{aligned} \quad (3)$$

Under (3) 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

$$\hat{T}_{\text{REG}} = \sum_{d=1}^5 \sum_{i=1}^{N_d} (\hat{\alpha}_d + \hat{\beta}_d X_{id}) \quad (4)$$

where $\hat{\alpha}_d$ and $\hat{\beta}_d$ are the OLS estimates of α_d and β_d in (3). Like the ratio estimator (2), (4) 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 of Population Estimation

To assess the performance of the three estimators of the population total described in Section 2.2 when the CCS design described in Section 2.1 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 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), leading to a total of 35 EDs (and hence 175 postcodes) being selected.

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
TOTAL	930	35

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) so that there is 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). Coverage in the CCS is set at approximately 90 per cent for households with 98 per cent of individuals within those households being counted. These can also be changed to see the impact of CCS response rates. Results in Brown *et al* (1999) show that, as one would expect, performance improves with increasing CCS response rates. In addition, robustness of the estimator with respect to dependence between the census and the CCS increases. 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. This is rather less than 1991 where overall coverage was around 98 per cent and aims to assess the robustness of the procedure to increased probabilities of underenumeration. The three estimators being evaluated are:

- 1) The ratio estimator with the DSE at the postcode level (Postcode Ratio)
- 2) The ratio estimator with the DSE at the HtC index level (Index DSE)
- 3) The regression estimator with the DSE at the postcode level (Postcode Regression)

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:

$$\text{RRMSE} = \frac{1}{\text{truth}} \times \sqrt{\frac{1}{1000} \times \sum_{j=1}^{1000} (\text{observed}_j - \text{truth})^2} \times 100 \quad (5)$$

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

$$\text{Relative Bias} = \frac{1}{\text{truth}} \times \frac{1}{1000} \times \sum_{j=1}^{1000} (\text{observed}_j - \text{truth}) \times 100 \quad (6)$$

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 three population estimators for the population total

	Postcode Ratio	Index DSE	Postcode Regression
Relative Bias (%)	0.27	0.26	0.41
RRMSE (%)	0.57	0.55	0.68

Table 2 summarises the results for the estimation of the total population by summing the individual age-sex estimates. There are not tremendous differences between the estimators. However, the two estimators based on the ratio model both have lower bias and lower total error. Looking at the total can hide problems with the estimation of the individual age-sex population totals. Figure 1 presents the results for the three estimators across the age groups.

Figure 1 clearly shows that across age-sex groups the three estimators are very similar in terms of RRMSE and always better than the census with the exception of men aged 85 years and over. With respect to bias the postcode regression estimator has a higher bias for the young age groups of both sexes. This is what gives the postcode regression estimator its high relative bias in Table 2. The two 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. It is also possible to plot the distribution of the individual errors from the 1000 CCSs for each age-sex group. This shows that while the estimators based on the ratio model have smaller inter-quartile ranges for the distributions of their errors they are slightly more prone to outliers.

This suggests that, in general, the estimators based on the ratio model are better with the Index DSE estimator being ‘best’ overall. However, this particular estimator 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 estimators 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.

There are also problems with both the ratio and the regression model as the census count gets small. 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. In the simulation about one third of the sampled postcodes counts are at the origin for most age-sex groups. The presence of the points at the origin tends to rotate the fitted line increasing the estimate of the slope leading to the positive bias. The origin points do not affect the ratio model as it is constrained to pass through the origin. However, 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. However, their affect is most dramatic in the oldest age groups when there are only a few non-zero census counts and the observed counts are in general all close to zero. It is this that generates the positive bias for these estimators that can be seen in Figure 1.

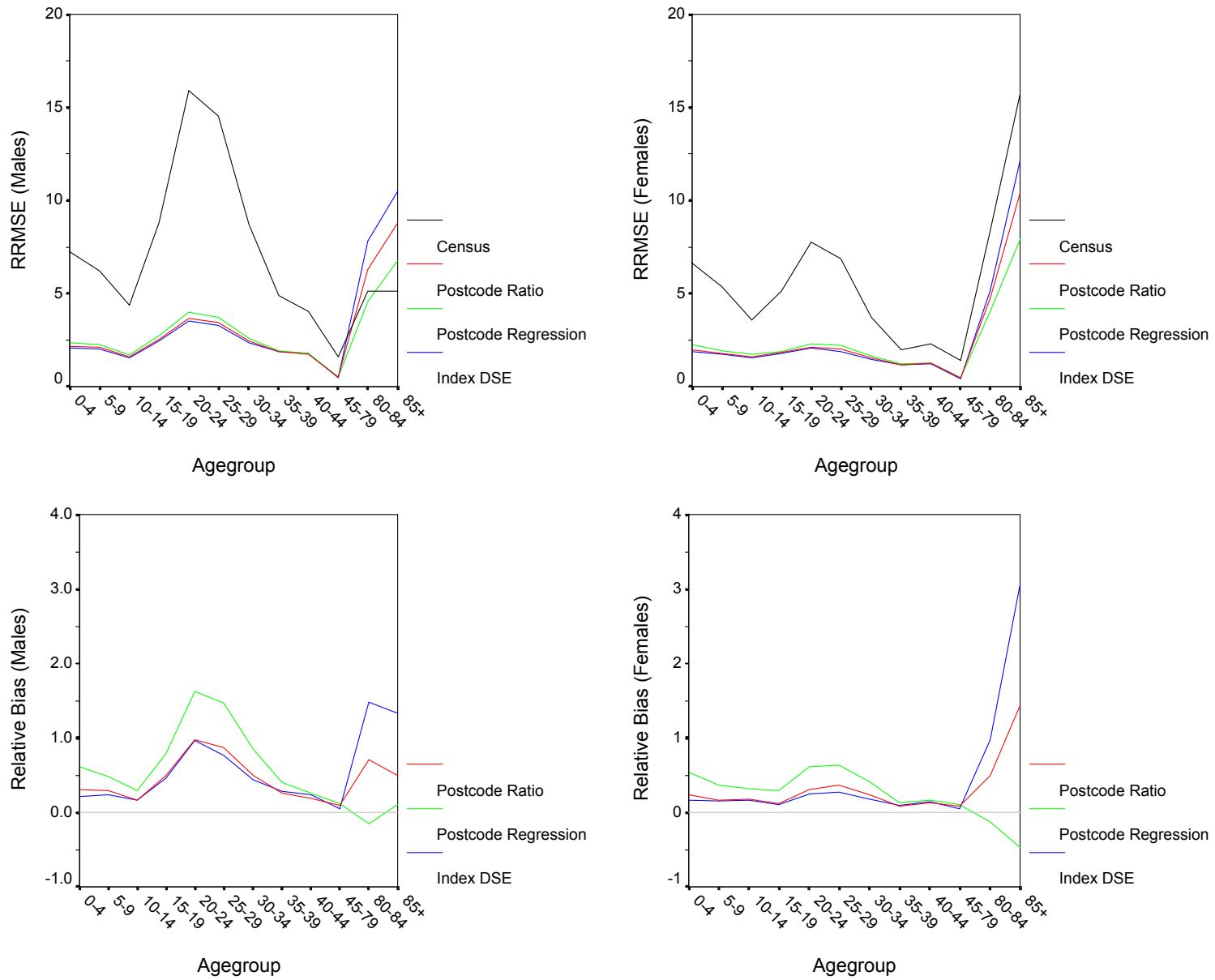


FIGURE 1: Performance of the three population estimators by age and sex compared to the census.

3.4) A Regression Model Robust to Zero Counts in the Census and CCS

The results in the previous section demonstrate a bias in the regression estimator. This arises when a high proportion of the postcode counts for a particular age-sex group are zero, making this point highly influential when fitting the regression line. The problem can be overcome by splitting the estimation into two parts; the first part estimates the true population count for those postcodes with a non-zero census count, the second part estimates the corresponding count for those postcodes with a zero census count. The model underlying this ‘mixture’ estimator within an age-sex group can be expressed as:

$$\begin{aligned}
 &\text{If } X_{id} > 0: \\
 &E\{Y_{id} | X_{id}\} = \alpha_d + \beta_d X_{id} \\
 &\text{Var}\{Y_{id} | X_{id}\} = \sigma_d^2 \\
 &\text{Cov}\{Y_{id}, Y_{jf} | X_{id}, X_{jf}\} = 0 \text{ for all } i \neq j
 \end{aligned} \tag{7}$$

$$\begin{aligned}
 &\text{If } X_{id} = 0 \\
 &Y_{id} = W_{id} (\mu_d + e_{id}) \text{ where } W_{id} \text{ takes the value one with probability } P_d \text{ and zero otherwise} \\
 &\text{and } e_{id} \text{ is distributed independently of } W_{id} \text{ with} \\
 &E\{e_{id}\} = 0 \\
 &\text{Var}\{e_{id}\} = \varpi_d^2 \\
 &\text{Cov}\{e_{id}, e_{je}\} = 0 \text{ for all } i \neq j \\
 &\text{leading to} \\
 &E\{Y_{id}\} = P_d \mu_d \\
 &\text{Var}\{Y_{id}\} = \mu_d P_d (1 - P_d) + P_d \varpi_d^2 \\
 &\text{Cov}\{Y_{id}, Y_{jf}\} = 0 \text{ for all } i \neq j
 \end{aligned} \tag{8}$$

In theory results from estimation based on the mixture model defined by (7) and (8) should be unbiased although possibly less efficient than (4), the standard regression estimator. However, early results show that in the simulation this is not the case. It turns out that the exclusion of the points with $X_{id} = 0$ makes the estimation of the regression relationship unstable in many cases as these account for a large proportion of the data. The regression relationship can be stabilised by collapsing strata defined by the HtC index. However, initial attempts at this have not reduced the bias as it is still difficult to estimate the parameters P_d and μ_d .

It is possible to apply this mixture model approach to the ratio estimator (2). The ratio model defined by (1) is robust to points at the origin, as the relationship is strict proportionality between the census and DSE counts. However, points for which W_{id} in (8) is equal to one introduce positive bias. This is easily shown as the exclusion of these points from the estimation turns the slight positive bias of the ratio estimator in Figure 1 to a negative bias. However, as with the regression model initial attempts in the simulation to estimate the parameters to adjust the ratio estimator have not been successful.

In practice, the results for the ratio model in Figure 1 show that overall it is more robust to the few points for which W_{id} equals one than the regression model is to many points at the origin. However, for a particular set of data it will not always be superior. Future work will focus on both improving the estimation of the parameters in (8) and specifying rules that the computer can use to make a decision on which estimator to use.

3.5) Matching Error and the Ratio Model

One of the most challenging aspects of using the DSE is the matching of the census and CCS databases to determine whether individuals have been counted twice or only counted once. Any error at this stage feeds directly into the DSE and becomes bias, the direction of which depends on the nature of the matching error.

Current research (Baxter, 1998) is developing an automated matching strategy that minimises the error from matching. It tackles the problem by first using exact computer matching, which is expected to handle the majority of cases. The remaining individuals are then matched using probability-based methods. Those records that cannot be matched this way will be matched by hand. The aim at the probability matching stage is to set the probability threshold high enough to reduce the risk of incorrectly matching two records. However, this is likely to increase the amount of hand matching that needs to be done and some individuals who should be matched will not be matched. When such matching error occurs it biases the DSE positively by decreasing the total number of people matched and therefore increasing the estimate of the population total through increasing the numbers in the off diagonal cells.

This type of matching error was introduced into the simulation programme described in Section 3.2. For each individual counted by both the census and the CCS a Bernoulli trial is used to establish whether they are correctly matched or not. The assumption is that the matching error is constant across all individuals. Once matching error is introduced the DSE estimator for Y_{id} may not be the most efficient. Two alternative estimators for the true population in a postcode are:

- Y_{MAX} = Maximum of the census and CCS counts for a postcode
- Y_{UNION} = Union of the census and CCS counts for a postcode

Any of the estimators of the true postcode population can then be used by ratio model (1) to estimate the population total. It is expected that they will each have the following different properties under the matching error that has been introduced into the simulations. Y_{MAX} will be robust to bias due to matching error but will be biased unless the CCS has a perfect response rate. Y_{UNION} will have a bias due to matching error and CCS non-response. However, it will have less non-response bias than Y_{MAX} and it should be more robust to matching error than Y_{DSE} .

For the postcode based ratio estimator the simulations have been re-run for three levels of matching error, 99.9 per cent of true matches are achieved, 99.5 per cent of true matches are achieved, and finally only 99 per cent of true matches are achieved. In each case the population totals have been estimated and these show the effect of this simple matching error on the estimators. Table 3 shows the results for the total population estimated by the postcode ratio estimator using the three different postcode counts.

TABLE 3
Estimation of the Overall Population Total in the Presence of Matching Error

	100% Match Rate			99% Match Rate		
	MAX	UNION	DSE	MAX	UNION	DSE
Relative Bias (%)	-1.45	-0.37	0.27	-1.45	0.47	1.71
RRMSE (%)	1.50	0.56	0.57	1.50	0.65	1.79

When matching is perfect Table 3 shows that there is little to choose between using the DSE count and the Union count when estimating the total population. The DSE has a slight positive bias, the reasons for which are discussed in the previous two sections. The estimator based on the Union count will also suffer from positive bias induced by using the ratio model. However, this is swamped by the negative bias due to taking no account of individuals missed by both. Once matching error is introduced the estimators based on both the Union and DSE counts now show a positive bias. This is what one would expect for the matching error that has been introduced. Table 3 now suggests the estimator based on the Union count is clearly better. However, this hides the fact that there are two biases, one from non-response and one from matching, which are going in opposite directions. If you consider the gross effect of the bias from the two causes the Union count has a relative bias of 1.21 per cent. This would feed into the calculation of RRMSE giving 1.98 per cent, worse than the DSE where all the biases are in the same direction. Table 3 also shows that for a good census and good CCS matching error very quickly makes the ‘simple’ maximum count approach a more efficient approach.

Looking at the aggregate level can hide what is going on for each individual agegroup. Figure 2 gives the results for the three estimators by agegroup for males. Figure 2 confirms that for a CCS with a good response rate there is not much to choose between the estimators in terms of RRMSE. The ratio estimator based on Y counts calculated using the DSE has a positive bias as seen in Figure 1. The other two estimators show negative bias due to people missed by both the census and the CCS. It is likely that this non-response bias should be greater but positive bias in the ratio estimation procedure is tending to cancel out its affect. The estimators have a better RRMSE than the census accept for the oldest two agegroups where the numbers being estimated are very small. The bias is always better than the census and although not presented here a similar pattern is observed for females.

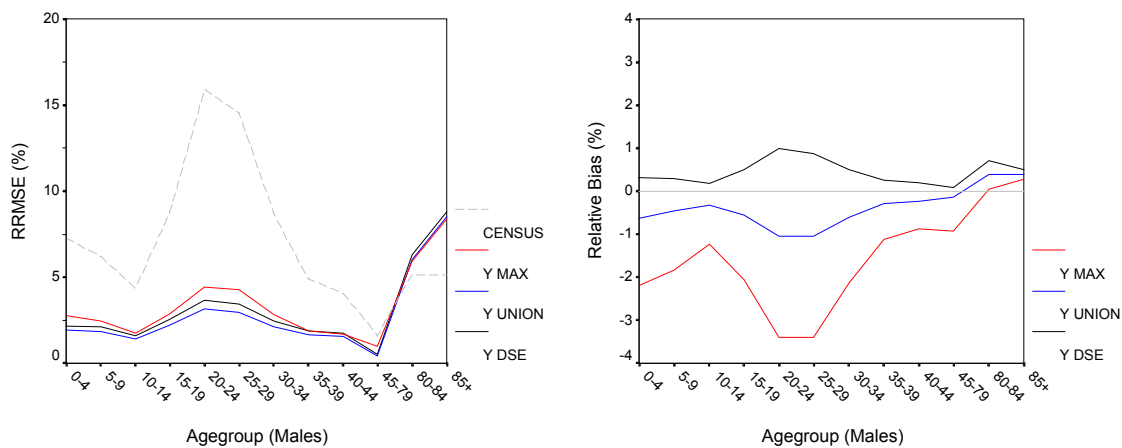


FIGURE 2: Performance of the different ‘Y Counts’ when there is no error due to matching.

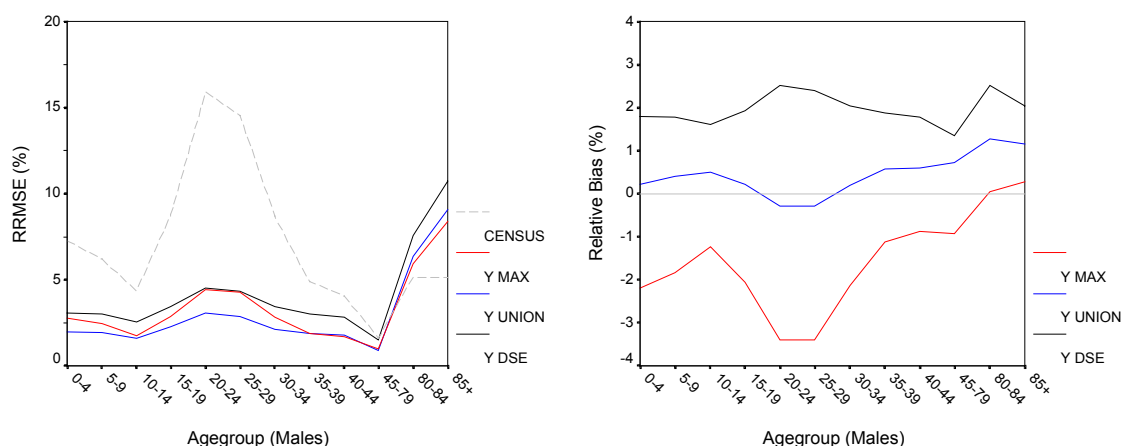


FIGURE 3: Performance of the different ‘Y Counts’ when matching success is 99 per cent.

For Figure 3, 99 per cent of the individuals counted in both the census and the CCS are matched. Figure 3 shows the same impact due to matching error on the estimation of individual agegroups as Table 3 demonstrates for the population total. The estimator based on the maximum count remains unaffected by the matching error while the DSE based estimator shows a clear increase in the positive bias and the union based estimator is now also tending to a positive bias. However, as pointed out earlier to accept the union based estimator as best relies on the dangerous assumption that two different biases will cancel each other. What this does clearly show is the impact of matching error and demonstrates the need to minimise errors at the matching stage in 2001. As the census and CCS response rates increase the bias in the maximum count due to non-response will tend to zero whereas bias due to matching error will always be present in the DSE and union counts. At this level of matching error the estimators still remain better than the census and as before the pattern for females is very similar.

3.6) Dependence and the Ratio Model

The simulations so far have assumed that the census and CCS are independent of each other, an assumption for the DSE to be unbiased. It is unlikely that this assumption will be satisfied in the reality. Brown *et al* (1999) extends the simulation study described in Section 3.2 to the case when an individuals probability of being counted in the CCS depends on whether they were counted in the census. This is introduced using the odds ratio between the census and the CCS. In this paper the simulations have been re-run for the following three values of the odds ratio:

- a) 1 (the independence case)
- b) 0.5 (people missed by the census are twice as likely to be counted by the CCS than those counted by the census)
- c) 2 (people counted by the census are twice as likely to be counted by the CCS than those missed by the census)

Odds ratios less than one represent the situation where the CCS finds the missed people while those who have participated in the census tend to refuse to respond to the survey. Odds ratios greater than one represent the situation where the CCS tends to miss the same people as the census. Table 4 presents the results from using the DSE with ratio estimation to estimate the total population for the different odds ratios. Figure 4 gives the results broken down by agegroup for males.

TABLE 4
Estimation of the Total Population for Different Levels of Dependence
Odds Ratio Between Census and CCS Counts

	0.5	1 (Independent)	2
Relative Bias (%)	0.41	0.27	0.04
RRMSE (%)	0.66	0.57	0.48

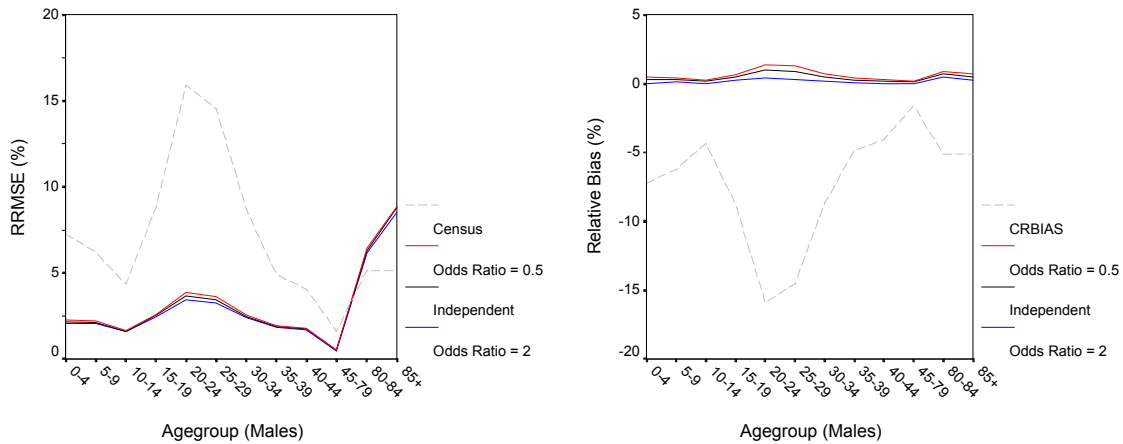


FIGURE 4: Population of the DSE Based Postcode Ratio Estimator Under Dependence

Both Table 4 and Figure 4 show that the estimator behaves as you would expect when the two counts are not independent. When the odds ratio is less than one this feeds a positive bias into the estimation, when it is less than one it feeds a negative bias into the estimation. However, Figure 4 also clearly shows that the estimation approach is robust to this level of dependence, a fact that is more clearly demonstrated in Brown *et al* (1999) for more extreme odds ratios and different response rates in the CCS.

3.7) Population Estimates for Variables Other Than Age and Sex

So far estimation has only considered partitioning the population by age and sex. Estimates by age and sex are key but estimates of the population by other characteristics are also of interest. The model specified in (1) or (3) can be used to make estimates for other census variables by applying it to postcode counts for those variables. The potential weakness is the fact that the probabilities of being counted by the census or by the CCS will be more heterogeneous as age and sex are the main predictors of non-response. For this reason in 2001 the age-sex totals will be considered the ‘gold standard’ and other estimates will be constrained to them. It is also possible to consider household variables as well. In this case the DSE is estimating the number of households by some variable of interest at the postcode level. The ratio or regression model is then applied to the postcode counts for households rather than individuals.

To demonstrate this, the same simulation has been used to estimate the adult population (those aged 16 and over) by primary activity last week as well as the number of households by tenure.

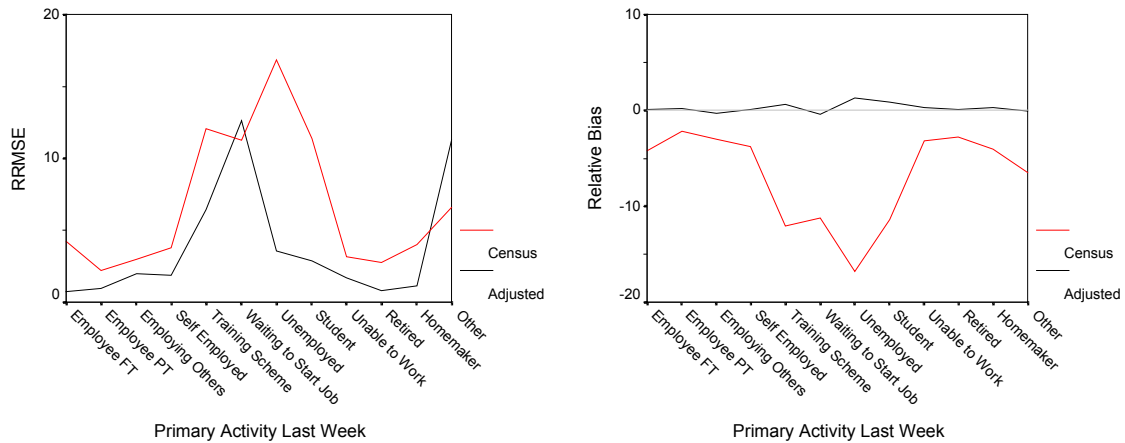


FIGURE 5: Performance of the regression based estimator for Primary Activity Last Week

Unlike a variable such as age, which is reasonably constant in terms of size across the population, the numbers in each category for primary activity last week vary quite dramatically. This is reflected in Figure 5 by the peaks in the RRMSE for certain categories with small populations to be estimated. However, the estimated count has, in general, a lower RRMSE than the census with a dramatic reduction in the bias and Figure 5 demonstrates the feasibility of estimating other individual characteristics.

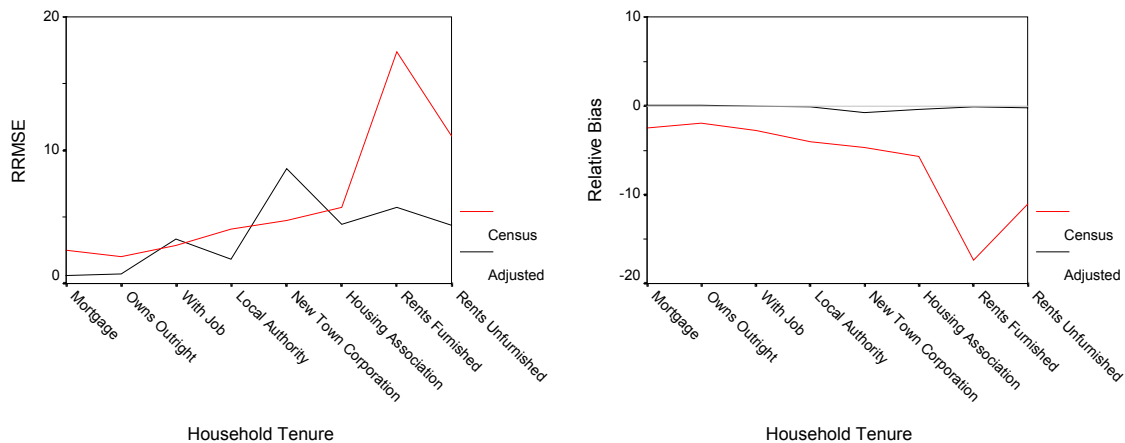


FIGURE 6: Performance of the regression based estimator for Household Tenure

Figure 6 again shows that the regression based estimator reduces the bias in the census, this time for estimates of the number of households by tenure. Again, the RRMSE is particularly large for one category, New Town Corporation, but there are only a small number of households in the simulation population of this tenure. For certain variables this effect could be reduced by collapsing some of the categories together when there is little evidence to suggest underenumeration would be differential across the categories. However, Figure 6 clearly demonstrates the feasibility of making household estimates.

6) Additional Considerations

6.1) Estimation of Overenumeration in the 2001 Census

All the work presented in this paper has assumed that overenumeration in the UK Censuses is minimal. In previous censuses no attempt has been made to adjust for overenumeration and it has been assumed that careful implementation of the census in the field minimises the risk of it occurring. It is very unlikely to occur by a person or household completing two forms for the same place but will happen when a person is erroneously added to a form for a household. Research is being carried out to assess its importance in the UK context. Methods of identifying overenumeration and erroneous census counts using the CCS and the matching exercise are also being investigated.

Given that erroneous counts and overenumeration can be identified in the CCS postcodes, this can be adjusted for when using the DSE to calculate the Y_{id} 's to put into the estimation models. It would not be necessary to adjust the census count when using it as the auxiliary variable in the ratio or regression models as the aim is an adjustment for net underenumeration, assuming that underenumeration is greater than overenumeration.

6.2) Variance Estimation

This paper has concentrated on the performance of the estimators in a simulation study where the truth is known. In 2001 the truth will not be known and therefore the measures of performance will be estimated variances for the estimators. Results for the postcode level regression estimator in Brown *et al* (1999) using the 'Ultimate Cluster Variance Estimator' were very promising. This estimator gave good coverage for estimated confidence intervals in the situation of independence between the census and CCS. However, more work is needed to assess both its robustness and stability.

7) Conclusions

This paper has presented research that is being undertaken as part of a research project by the Office for National Statistics to estimate and adjust for underenumeration in the 2001 Censuses of the UK. The standard technique for estimating underenumeration, dual system estimation, has been combined with both ratio and regression estimation models. The simulation study, which extends Brown *et al* (1999) to include heterogeneity in the CCS as well as the census, shows that while estimators based on both models perform well there is a robustness issue. This particularly affects the estimator based on the regression model. A 'mixed' model is suggested as a more robust alternative to the simple regression model. However, initial simulation results suggest that this is a difficult to estimate alternative model.

The simulation study has also been extended to include a simple form of matching error. As expected, the dual system estimator translates errors of the sort introduced into bias. There are alternative estimators to the dual system estimator. However, much more work is needed before levels in matching error could be set where one of the alternatives would be chosen in preference to the dual system estimator. A simple extension also demonstrates the robustness of the approach to some dependence between the census and the CCS.

The paper has mainly concentrated on the estimation of the population by age and sex. While this is the key concern estimation by other variables will also be required if a full adjustment of the census database is to take place. Simulation results in this paper clearly demonstrate that the methods proposed for estimation by age and sex can also be applied to estimation by other individual and household characteristics.

The work presented in this paper is ongoing. The next major steps will be the further development of robust estimation models for both the regression and ratio approaches. More work is also needed to more fully assess the effect of increasing heterogeneity on the estimator. Issues relating to overenumeration, and its estimation, along with variance estimation also need to be fully addressed to ensure that a robust and efficient estimation strategy is adopted in 2001.

References

- Alho, J. M., Mulry, M. H., Wurdeman, K. and Kim, J. (1993) Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, **88**, 1130-1136.
- Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics* **10**, 245 - 256.
- Baxter, J. (1998) One Number Census Matching. *One Number Census Steering Committee Working Papers ONC(SC)98/14*. Available from the Office for National Statistics, Titchfield. (jennet.baxter@ons.gov.uk)
- Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999) A methodological strategy for a One Number Census. To appear in *Journal of the Royal Statistical Society A* **162**, Part 2.
- Diamond, I. D. (1993) Where and who are the 'missing million'? Measuring census of population undercount. In *Statistics Users' Council Conference Proceedings on Regional and Local Statistics, 16th November 1993*. Published by IMAC Research, Esher.
- 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.
- 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.
- 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.