



ONS(ONC(SC))98/05

ONE NUMBER CENSUS STEERING COMMITTEE

A Methodological strategy for a One Number Census in the United Kingdom

1. This paper has been submitted to the Royal Statistical Society Series A for publication. It is based on the presentation given at the Cathie Marsh Memorial Seminar by Ian Diamond and Andy Teague.
2. The majority of the paper is based on the papers for the last ONC Steering Committee meeting. However, Section 4 describes the use of a combined Dual System and regression estimator approach instead of the regression estimator previously described.
3. **The Steering Committee are asked to:**
 - a) **note the paper and the progress made;**
 - b) **provide any comments at the meeting on the 27 April 1998, or in writing by 10 May 1998.**

**Lisa Buckner
Census Division
Office for National Statistics**

**Room 4200W
Segensworth Road
Titchfield
Fareham
HANTS
PO15 5RR**

March 1998

**A METHODOLOGICAL STRATEGY FOR A ONE NUMBER CENSUS
IN THE UNITED KINGDOM**

**By J. J. Brown⁽²⁾, L. J. Buckner⁽¹⁾, I. D. Diamond⁽²⁾,
R. L. Chambers⁽²⁾, A. D. Teague⁽¹⁾**

For correspondence:

Lisa Buckner

**⁽¹⁾ Office for National Statistics
Room 4200W, Census Division
Segensworth Road
Titchfield
Fareham
Hampshire
PO15 5RR**

**⁽²⁾ University of Southampton
Department of Social Statistics**

**Tel: 01329 813507
Fax: 01329 813532
Email: lisa.buckner@ons.gov.uk**

SUMMARY

As a result of lessons learnt from the 1991 Census, a research programme was set up to seek improvements in census methodology. Underenumeration has been placed top of the agenda in this programme, and every effort is being made to achieve as high a coverage as possible. In recognition, however, that 100 per cent coverage will never be achieved, the One Number Census (ONC) project was established to measure the degree of underenumeration in the 2001 Census and, if possible, fully adjust the outputs from the census for that undercount. This paper describes the background to the project. It should be emphasised that the research is still being carried out and as such, the methodology proposed is subject to refinement. These refinements to the methodology will be presented in a series of consultation papers.

Keywords: CENSUS UNDERENUMERATION; CENSUS COVERAGE SURVEY; SAMPLING; REGRESSION ESTIMATION

1. INTRODUCTION

One of the major uses of censuses in the UK is in providing figures on which to rebase the annual estimates of the population by age and sex. This base needs to take into account the level of underenumeration in the census, which has traditionally been measured by the use of a post-enumeration survey (PES) and through comparison with the estimate of the population based on the previous census. Until the 1991 Census, there was close agreement between the adjusted census count (census + PES) and the estimate based on the previous census. Moreover, the estimated level of underenumeration was relatively small (less than one per cent). In 1991, the level of underenumeration was much higher (2.2 per cent); underenumeration did not occur uniformly across all socio-demographic groups and parts of the country (for example, it was estimated to be over 20 per cent for young males in inner cities); and there was a significant difference between the survey-based estimate and that rolled forward from the previous census. This led to difficulties for the Census Offices of England and Wales, Scotland and Northern Ireland in rebasing the population estimates as well as for the census users in interpreting census counts.

The 1991 Census Validation Survey (CVS) - as the post-enumeration survey was known - did not adequately identify the extent and distribution of underenumeration. The CVS suggested a total of 290 thousand people were missed by the census in England, Wales and Scotland, whereas demographic estimates of the population (the estimate of the population based on the 1981 Census - births minus deaths plus net migration in the intercensal period) indicated a figure of around 1.2 million people were missed. It was decided, on balance, that at the national level, the 1981-based demographic estimate was more reliable (OPCS 1993, 1994). Several differing population counts were therefore available for 1991: the unadjusted census count; the adjusted census count; and the demographic estimates. This made the distribution of the undercount to local areas difficult and caused considerable confusion amongst customers about how the differences between the rebased population estimates and the census counts should be interpreted.

The priority for the development of the 2001 Census is to ensure that the maximum possible coverage is achieved, and in particular that the differential nature of any underenumeration is

minimised. To this end, the methodology for carrying out the Census is being reassessed so as to reduce the burden on the public and to use resources to their best effect.

Despite efforts to maximise census coverage, it is only realistic to expect there to be some degree of underenumeration. The One Number Census (ONC) project aims to measure this level of underenumeration in the most acceptable way, to provide a much clearer link between the census counts and the population estimates, and if possible to adjust all the census counts (which means the individual level database itself) for underenumeration. All counts will then add to 'One Number'. This has entailed a re-think of the design of the post enumeration survey and how this should be integrated with other measures of underenumeration provided by administrative records and demographic analysis. Note that many of the references to the 1991 Census refer to the censuses carried out in England and Wales, and in Scotland. In contrast, the methodology for a ONC in 2001 is being developed for application in all four countries of the United Kingdom.

2. OVERVIEW OF THE ONC METHODOLOGY

The purpose of this paper is to make public the research that has already been undertaken towards a ONC in 2001 as presented at the 1997 Royal Statistical Society Cathie Marsh Memorial Lecture. The process outlined below represents the proposed methodology for a ONC. However, by its nature as a development project, the methodology described here is subject to change. The ONC process can best be considered as consisting of the four main stages summarised below and illustrated in Figure 1. The inputs to the process are discussed in more detail in Sections 3 to 5.

(Figure 1 here)

The first two stages will be to produce the best estimate of the population by age and sex at national and subnational levels (average of one and a half million people). The extension of the methodology to give estimates at the local authority district level is presented in ONS (1998). These stages are essential for rebasing the population estimates. The third stage will produce estimates for lower levels of geography and for other characteristics of people and households. The final stage is either to impute records for households that are estimated to have been missed and people estimated to have been missed from counted households or to produce a database weighted for underenumeration. This last stage would allow all statistics based on the 2001 Census to aggregate to 'One Number'. These last two stages have less relevance for central government distribution of resources but are relevant for local authorities who need to allocate resources at a local level. In addition, it is of course important that users know the implications of any undercount on figures for small population groups and geographical areas.

2.1 Stages 1 and 2 - National and Subnational level estimates

To estimate the population by age and sex at the subnational level, counts from the 2001 Census will be adjusted for estimated net underenumeration using a post-enumeration survey to be known as the Census Coverage Survey (CCS). These subnational level estimates will then be aggregated to produce a national census-based estimate and compared with a demographic estimate of the national population rolled forward from the

previous census. Charlton *et al.* (1997) summarise work underway to optimise the methodology used to produce the demographic estimates.

The fieldwork for the CCS based on the re-enumeration of a sample of whole postcode units, stratified by county and an index based on how difficult the postcode is expected to be to enumerate, was piloted in the Brent area of London following the 1997 Census Test. A short questionnaire was used to collect information on characteristics believed to be associated with underenumeration. The simplicity of the questionnaire and the fact that sampling whole postcodes makes efficient use of interviewer time, makes a much larger sample size possible than was the case for the 1991 Census Validation Survey. It is, however, inevitable that the CCS will fail to find all the missing people and the possibility of using ancillary information based on administrative records has been investigated.

2.2 Stages 3 and 4 - Small area estimation and imputation

The production of adjusted census counts for small areas (and ultimately an adjusted census database) represent the final goals of the ONC process. Models developed on the sampled postcodes in the CCS linking the observed census characteristics and the estimated missing population will be used to predict the number missed in non-sampled postcodes. These models will be used to estimate the number of people missed in enumerated households and those in wholly missed households for each postcode. As stated above, the precise method for creating individual records and allocating them to household units has not yet been developed. However, possible approaches for this are discussed in Section 5.

3. CENSUS COVERAGE SURVEY

The aim of the CCS following the 2001 Census is to facilitate the estimation of underenumeration at a subnational level (by age and sex); and to allocate this underenumeration down to small areas. The precise unit of aggregation has still to be agreed but the design described below uses counties or groups of counties with an approximate total population of one and a half million. The design framework does not rely on this choice of aggregation. Changing the level of aggregation only has implications for the sample size and achieved accuracy. A simulation study is undertaken to assess the design and direct estimation procedures.

Following the 1991 UK Census a Census Validation Survey (CVS) was carried out in England, Scotland, and Wales. The survey aimed to estimate net underenumeration and to validate the quality of census data. The second of these aims required the re-enumeration of a sample using the entire census form. This requirement is costly, due to the time required to fill out this form, resulting in a small sample size. It is proposed that the survey in 2001 should address coverage exclusively. Information on the quality of census data would be obtained from the question testing programme, the 1997 Census Test and possibly through a survey carried out after the Census Dress Rehearsal in 1999. This allows for a much shorter doorstep questionnaire. Savings in time can be translated into a larger sample size.

The proposal is for a postcode-unit based survey. This requires the re-enumeration of a sample of postcode units rather than households. This clustering also helps to enable a

larger sample size. While that does not necessarily improve the direct estimation of underenumeration due to the increase in variance as a result of correlation between households within postcodes, it is important for estimating adjustments at lower levels.

The sample of postcodes needs to be sufficiently large to estimate the total population by 24 age-sex groups, for each design level group¹. At the design level, postcodes are stratified into groups by a ‘Hard to Count’ (HtC) index and then size. It is expected that underenumeration will, at a local level, be higher in certain areas characterised by particular social, economic and demographic characteristics. For example, it is known that people in dwellings occupied by more than one household (multi-occupancy), will have a relatively high probability of not being enumerated. Therefore, a national HtC index was formed for 1991 Census enumeration districts by ranking the enumeration districts with respect to a series of variables and then assigning normal scores based on those ranks. The following variables were used:

- percentage of heads of household who experienced language difficulty as defined by country of birth;
- percentage of young people who migrated in to the enumeration district in the last year;
- percentage of imputed residents for the enumeration district;
- percentage of households which lived in multiply-occupied buildings; and
- percentage of households which were private rented.

At a national level these were divided into quintiles with each quintile assigned a value from 1 (easiest to count) to 5 (hardest to count). The components of the HtC index were chosen to represent characteristics found to be important after the 1991 Census by the Office for National Statistics (ONS) and the Estimating With Confidence Project (Simpson *et al.*, 1997). The problem is to estimate the 24 age-sex totals such that each has an expected relative standard error (RSE)² of less than α per cent where α is chosen depending on the required accuracy and cost constraints.

In general, postcode level information, beyond number of addresses, is not known. This leads to a two-stage design, selecting enumeration districts as Primary Sampling Units (PSUs) and then sampling postcodes as Secondary Sampling Units (SSUs) within selected enumeration districts. Clustering from the two-stage design has cost advantages for a fixed number of postcodes but efficiency disadvantages when the characteristics of postcodes are positively correlated within enumeration districts.

In order to make direct estimates from the CCS the quantities of interest are:

Z_{aiedc} = 1991 adjusted census count for age-sex group a of postcode i , within enumeration district e , in HtC category d of design level group c .

X_{aiedc} = 2001 unadjusted census count.

Y_{aiedc} = “True” 2001 count (given by the CCS for those postcodes in sample).

where:

¹ Each design level group is either a single county or group of smaller contiguous counties. The age-sex groups to be estimated are: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-79, 80-84, 85+ for males and females.

² RSE (also called coefficient of variation) = $\frac{\sqrt{\text{var}(T)}}{T} \times 100$

$c = 1 \dots C$ design level county groups in England & Wales.
 $d = 1 \dots 5$ HtC categories of postcodes.
 $e = 1 \dots M_{dc}$ enumeration districts in HtC category d of group c .
 $i = 1 \dots N_{dc}$ postcodes in HtC category d of group c of which n_{dc} are in the sample S_{dc} , the rest are in the non-sample R_{dc} .
 $a = 1 \dots 24$ age-sex groups (0-4, 5-9, 10-14, ..., 40-44, 45-79, 80-84, 85+).

For direct estimation from the CCS it is required that the total population counts by age-sex and design level group, given by T_{ac} , be estimated to a certain degree of accuracy. This is treated as 24 similar estimations within each design level group. For this reason the design and estimation for one age-sex by design level group is described below. The same methodology applies for all other age-sex groups and in the following the subscripts a and c are dropped.

3.1 Stage One of the CCS design

A robust approach to design for stage one of the CCS assumes a stratified homogeneous super-population model for the distribution of true 2001 counts within enumeration districts with simple random sampling within each stratum. Within a design level group the enumeration districts are stratified by the HtC index. This is important as, within the design group, undercount will depend on the characteristics of the PSUs. It also ensures that the CCS sample is spread across the full range of enumeration districts. Further stratification by size based on the 1991 adjusted census counts improves efficiency by reducing within stratum variance. Ideally one would like to use the 2001 unadjusted counts but the CCS must be ready for the field directly after the census so this is not possible. It is expected that the final design will use 1991 based projections of the population in 2001.

Allowing for $h = 1 \dots H_d$ size strata within each HtC category the model for a given age-sex group within a design level group can be written as:

$$\left. \begin{aligned} E\{Y_{ehd}\} &= \mu_{hd} \\ \text{Var}\{Y_{ehd}\} &= \sigma_{hd}^2 \end{aligned} \right\} e \in h \text{ within } d \tag{1}$$

$$\text{Cov}\{Y_e, Y_f\} = 0 \text{ for all } e \neq f$$

Assuming no second stage sample, estimation of the required total is straightforward under the model in (1) using a stratum by stratum expansion estimator. From this it is possible to calculate the number of enumeration districts that need to be sampled if there was no second stage sample. However a second stage of sampling within selected enumeration districts is proposed and so a regression estimator will be used to compensate for the resulting loss in efficiency. The practicalities of choosing stratum boundaries and allocation to strata are discussed in Section 3.4.

3.2 Stage Two of the CCS design

The second stage of the CCS design consists of a random selection of postcodes within selected enumeration districts. The design is based on a selection of the same number of postcodes within each sampled enumeration district using simple random sampling without replacement. Given that size stratification and optimal allocation was used at stage

one of the design, so that probability of selection of an enumeration district is approximately proportional to its size, this means that within a HtC category each postcode has approximately the same probability of inclusion in the sample.

3.3 The CCS model for estimation

It is sensible to assume that the 2001 Census count and the CCS count within each postcode will be related. If this is not true then suspicion should fall on one of the counts. Further, a linear regression relationship between the two counts may well be appropriate, with the possibility of a non-zero intercept. This term is needed as in some postcodes the census can miss all the people from a certain age-sex group. Given that we know from the 1991 Census that age and sex are crucial to undercount, as well as local characteristics, it is sensible for each design level group to consider a model within age-sex groups for each HtC category. The simple regression model stratified by HtC index for an age-sex group is:

$$\left. \begin{aligned} E\{Y_{id}|X_{id}\} &= \alpha_d + \beta_d X_{id} \\ \text{Var}\{Y_{id}|X_{id}\} &= \sigma_d^2 \end{aligned} \right\} i \in d \quad (2)$$

$$\text{Cov}\{Y_i, Y_j | X_i, X_j\} = 0 \quad \text{for all } i \neq j$$

Substituting the ordinary least squares (OLS) estimators for α_d and β_d into (2), and remembering that this is a model within age-sex by design level group, it is straightforward to show (Royall, 1970) that under this model the Best Linear Unbiased Predictor (BLUP) for the population of interest's overall total T is:

$$\hat{T} = \sum_{d=1}^5 \left\{ T_{Sd} + \sum_{i \in R_d} (\hat{\alpha}_d + \hat{\beta}_d X_{id}) \right\} \quad (3)$$

where T_{Sd} is the total for sampled postcodes in category d of the HtC index and R_d is the set of non-sampled postcodes in category d of the HtC index. Strictly speaking the model specified by (2) is known to be wrong. The covariance assumption in the regression model ignores the fact that postcode counts are correlated within enumeration districts by the design. However, the simple two stage model proposed by Scott and Holt (1982), which assumes independence between PSUs, is still reasonable. Under this model Scott and Holt (1982) state that the OLS approach remains unbiased, and therefore (3), with only a small loss of efficiency.

The variance of $\hat{T} - T$, the estimation error associated with (3), can be estimated using the model given by (2). Unlike (3), this is sensitive to mis-specification of the variance structure even when the design is *approximately* balanced with respect to the auxiliary variable (Royall and Cumberland, 1978). Consequently, as the postcodes are clustered within enumeration districts, it is proposed that the conservative ultimate cluster variance estimator, a variant of the random groups approach, be used. Once the variances are estimated an estimated RSE can be calculated for each age-sex group total.

3.4 Case Study: A Prototype Stage One CCS Design for Hampshire

Hampshire was chosen purely for convenience to examine the feasibility of Stage One of the design. It was considered to be an ‘average’ county with just over 3,000 enumeration districts and includes two middle-sized cities. Some counties are considerably smaller hence the need in some cases to group contiguous counties at the design level.

The first stage of the simulation was to calculate a national HtC index using those enumeration districts with a non zero population in the 1991 Census. Within Hampshire there are 3,305 enumeration districts of which 3,229 had a non zero population in the 1991 Census and therefore a HtC index value. The distribution of the districts by the index is given in Table 1 and reflects the presence of the cities of Portsmouth and Southampton, with a predominance of enumeration districts, and therefore postcodes, in the harder to count categories.

(Table 1 here)

Within a HtC by design level group, estimation is required for each age-sex group. Consequently there are 24 potential size variables, the Z_{ae} 's, which can be used for stratification. The solution adopted here is based on a multivariate approach that uses six key age-sex groups, males and female 0-4, males 20-24, males 25-29, males 30-34, and females 85+. The choice of these key variables is based on a coverage analysis of the 1991 Census. In addition, 28 large enumeration districts with very high counts for males in the ages 20-34, were included in the final design with probability one. On the remaining 3201 districts principal components analysis was used to reduce the number of size variables. The first three component scores defined by these key size variables, which accounted for over 96 per cent of their original variability, were then used within each HtC index category to form strata by applying Ward's linkage (SAS Institute Inc., 1990) in a cluster analysis. A minimum cluster size of at least two enumeration districts was imposed. Clusters based on single enumeration districts were highlighted as outlying and included in the sample with probability one.

A design variable W_e based on the chosen principal components was then constructed as follows:

$$W_e = \frac{|V| \times \sum_{j=1}^3 P_{je}}{\left\{ \sum_{j=1}^3 \text{var}(P_{je}) \right\}^{1/2}} \quad (4)$$

where P_{je} is the j^{th} component score for the e^{th} enumeration district, and V is the variance-covariance matrix of the six original size variables calculated from the 3201 enumeration districts used in the principal component analysis. Using the determinant of this matrix as a measure of variability in the original data, and bearing in mind that principal components are orthogonal, the variance of the design variable in (4) is therefore equal to the original variability of all six size variables. The design variable W_e was then used to calculate the

total sample required to estimate its population total with an RSE of one per cent³. Neyman allocation was used to allocate this sample to the strata with the condition that the minimum stratum sample was one enumeration district.

Several different size stratifications were tried by varying the number of clusters formed in the clustering algorithm. In general, increasing the number of clusters brings down the total sample size as there is less within cluster heterogeneity. However, as clusters become more homogeneous the number of single enumeration district clusters identified by the algorithm increases. Furthermore, this increased homogeneity results in an increased number of optimal strata samples of less than one. The final design and allocation is given in Table 2.

(Table 2 here)

From Table 2 it would appear that more size strata would further reduce the sample but the gains are small and these are countered by more enumeration districts being allocated to clusters of size one. This increasing of outlying enumeration districts as a result of requiring more clusters may be reduced by applying other clustering algorithms in the final design. Further work to identify the characteristics of these outlying enumeration districts will also be necessary when the final design is calculated for all design level groups.

The design in Table 2 gives a total first stage sample of 347 enumeration districts, approximately a 10 per cent sampling fraction. To assess how well the design works for each individual Z_{ae} , rather than W_e , the expected RSEs were calculated for the 3190 enumeration districts not classified as outlying and taking a sample of 308. These ranged from 1.4 per cent for those males aged 0-4 to 4.6 per cent for those males aged 85+. The six age groups in the design variable all had expected RSEs of less than 1.7 per cent.

The design proposed for the first stage is standard. The auxiliary information is used to stratify, a standard procedure in both the model-based and design-based frameworks for making efficiency gains. The estimation model is chosen to make further efficiency gains using the additional auxiliary information available from the 2001 Census. These gains are related to the variability in census coverage as this affects the conditional variance in the model. For this reason giving more weight to the hardest to count categories is being investigated as these are expected to have more variable census coverage. However, the conditional variance will always be less than the marginal variance when a regression model is sensible, leading to some efficiency gain and introducing weighting by HtC category is not expected to change the overall sample requirement. The case study for Hampshire deals with the practical application of the design. It shows that the theoretical framework proposed can be applied to an actual county with feasible results. However, Table 2 does not represent the final design for the 2001 CCS in Hampshire. In the final design it is likely that the Isle of Wight will be included in a group with Hampshire.

³ An RSE of one per cent for total T translates into an approximate 95 per cent confidence interval on T of ± 2 per cent.

3.5 Extension to National Sample Size

Given the design described above it is necessary to estimate a sample size for the national sample (to cover England and Wales) for a range of design RSEs. Counties are quite variable in the number of enumeration districts they contain. However, the heterogeneity amongst enumeration district population counts within counties does not vary to the same extent. For this reason it is proposed that contiguous counties are grouped to make pseudo counties, similar in size to Hampshire, of about 3000 enumeration districts or approximately one and a half million people. An initial grouping has been made which reduces the 55 England and Wales counties to 34 groups. This grouping also accounts for splitting Inner London, Outer London, Greater Manchester, and West Midlands as these are much larger than 3000 enumeration districts.

The design has been implemented in Kent and West Yorkshire for a range of RSEs. These counties were chosen as Kent is approximately 3,000 enumeration districts, the average size required, and West Yorkshire is the largest single county which has not been split. The results are in Table 3.

(Table 3 here)

These two counties cover 7,256 enumeration districts out of approximately 110,000. Using linear extrapolation it is possible to extrapolate to national sample sizes and get approximate figures of:

- 40,000 postcodes (approximately 600,000 households) for an RSE of 1.0 per cent.
- 25,000 postcodes (approximately 375,000 households) for an RSE of 1.5 per cent.
- 19,000 postcodes (approximately 245,000 households) for an RSE of 2.0 per cent.

This translates to sampling between 2.5% and 1% of all households in England and Wales.

This simple extrapolation very much depends, of course, on Kent and West Yorkshire being a good representation of all design level groups. It should also be noted that the precise formulation of the HtC index is still under development.

3.6 Simulation Study of the CCS Design

The aim of the simulation study is to examine the performance of the CCS design, and particularly the gain from regression estimation, when the second stage sample is taken. Anonymised individual records from the 1991 Census, augmented by the HtC index, for one complete district from a county in England and Wales were used in the simulation. The district is treated as a design level group and has 450,000 individuals within 170,000 households. It consists of 11,000 postcodes (141 with only one person and 46 with over 200 people) and 900 enumeration districts (five have only one postcode, one has 40 postcodes, and the median is 14 postcodes)⁴. The distribution of enumeration districts by HtC index is given in Table 4.

(Table 4 here)

⁴ The numbers given are approximate for confidentiality reasons.

The distribution in Table 4 is reasonably uniform. This is important as it is necessary to avoid extremes, especially a situation where the easiest to count group dominates as this would tend to make the overall performance of the design too optimistic.

Treating these census records as corresponding to a real unobservable population, the first step of the simulation was to create a census. Each individual was given a fixed probability of being counted in a census based on their age, sex, and enumeration district HtC index. This was done by simple random sampling with replacement from the population of Estimating With Confidence enumeration district adjustment factors. These are the 'best guess' at small area coverage for the 1991 Census. To create a census, an independent Bernoulli trial was carried-out for each individual. Certain rules were then applied to ensure that counted households had a sensible structure. Households were excluded if:

- any children aged 5-15 were missed from a counted household
- all household members aged 16 and over were missed
- one partner from an elderly couple was missed.

This strategy for excluding households is not a perfect representation of reality as the rules do not cover all possible scenarios. Its advantage is simplicity as it produces missed households without the need to simulate dependence in the Bernoulli trials such that the probability of an individual going missing, given that other members of their household were missed, would increase.

For the CCS, the design procedure used for Hampshire (see Section 3.4) was followed but based on an RSE of 2.5 per cent to reflect the smaller population of PSUs. The final design and allocation is given in Table 5. The design in Table 5 was fixed throughout the simulation and used to get a total sample of 85 enumeration districts. A fixed sample of four postcodes (or the number of postcodes in the enumeration district if less than four) was taken at the second stage. For each sample the totals for each age sex group were estimated, the variances calculated using the ultimate cluster variance estimator and estimated RSEs calculated. Ideally, it would be desirable to simulate one CCS per census as this most accurately reflects real life. Computationally, censuses are time consuming to simulate so a compromise of 10 CCSs for each of 100 censuses was adopted.

(Table 5 here)

(Table 6 here)

Table 6 shows that the procedure does well on average and in all cases the average estimated RSE is better than the RSE one would expect to get if the stratified expansion estimator (used in the design) was applied with no second stage sample. This shows that on average the regression estimator has enough extra efficiency over the stratified expansion estimator to recover the loss of efficiency due to two stage sampling. It is also able to reduce the RSE in those age groups not included in the clustering to produce size strata and the construction of the design variable W_e . However, the standard errors do show that for most age-sex groups it cannot be guaranteed that the regression estimator will do better for every CCS. In those instances applying the stratified expansion estimator may be more efficient.

The simulation shows that for a perfect CCS the proposed design in conjunction with the regression estimator performs well. In Section 4 the more realistic situation of dependence

between the census and CCS with CCS non-response is examined in detail for the regression estimator model.

4. DEPENDENCE, NON-RESPONSE AND UNDERENUMERATION ESTIMATION METHODS

The theory underlying use of the regression estimator in the CCS design assumes that the CCS count is perfect for the sampled postcodes. This can be extended to allow for some non-response in the CCS by assuming that between the census and the CCS there is a complete count and no persons are missing from both. In this case the regression estimator uses the union of the census and CCS counts as its 'Y' count while still using the raw census count as the auxiliary variable. Clearly one would expect the regression estimator using this union count to have a negative bias if people are missing from both counts, assuming that the regression model underlying this estimator is appropriate.

The assumption that no one is missing from both counts effectively requires dependence between the census and CCS, as the CCS must find the people that the census missed. It is unlikely that the CCS will be able to find all the missed people and there are estimators that try to account for this. One well-known approach used by the US Census Bureau is known as the Dual System Estimator (DSE). Hogan (1993) covers the implementation of this methodology as it was applied to the 1990 Census in the US⁵. The DSE assumes that the census and CCS counts are independent and when this assumption holds, the DSE gives an unbiased estimate of the total population. There is another assumption, that of homogeneous capture probabilities. The US Bureau try to approximate this by forming post strata based on the characteristics which cause the most heterogeneity in the capture probabilities. As with the union count the DSE count for a sampled postcode can be used as the 'Y' in the regression estimator to adjust for people missed by both counts. As a postcode is a small population in a generally small geographic area, with the counts split by age and sex, the homogeneity assumption is expected not to be seriously violated. In the situation where people missed by the census have a higher chance of being missed by the CCS than those counted by the census, one would expect the DSE count regression estimator to still under estimate but not by as much as the union count regression estimator. When the reverse happens and the CCS is very good at finding the missed people (the requirement for getting unbiased estimates when using the union count in the regression estimator) one would expect the DSE count regression estimator to over-estimate.

One solution to the problem of dependence between the census and the CCS is to extend the DSE to a Triple System Estimator (TSE). This requires a third list to which both the census and CCS counts can be matched, and has been investigated by Zaslavsky and Wolfgang (1993) and Darroch *et al.* (1993). The advantage of having the third list is that it is then possible to estimate two-way interactions between counts derived from the different sources and the independence assumption is no longer necessary. This can be seen as equivalent to a three-way contingency table model, with no three-way interaction term, and can be used to calculate an estimate for the missing cell count of people who appear on none of the three lists. In theory this triple system approach is superior to a dual system

⁵ This is part of a special section 'Undercount in the 1990 Census' in volume 88 of JASA. This section contains several papers on the practical and theoretical aspects of using DSE methodology.

approach but it requires a third source of data, typically from administrative sources. Obtaining a good third list that has reasonable coverage with no over coverage and that correctly locates people is not straightforward.

A common problem for all the above methods is the ability to match individuals between the various lists. Mis-matching can be a major source of bias either by creating too many matches (negative bias) or not enough (positive bias). This problem is accentuated in the case of three lists where the data on the third list may have been collected for other purposes and may not have information which could be used for matching. This is an area of recent development, an example is Kendrick (1997), and it is intended to draw on knowledge gained from this expertise to develop a matching procedure for use in the ONC project.

Section 4.1 describes simulations which investigate dependence between the census and the CCS for different ‘Y’ counts in the regression estimator. In Section 4.2 the additional effect of list inflation on the third list used in the TSE is also considered.

4.1 Simulation study of the impact of correlated non-response on the CCS design

The simulation described in Section 3.6 did not consider the problem of non-response in the CCS or dependence between the census and CCS counts. In reality both of these situations are likely. To investigate the impact of this correlated non-response the simulation program used in Section 3.6 was extended. Dependence between the census and CCS was achieved using a method which involves varying the odds ratio for the probability of being counted by the CCS relative to the probability of being counted by the census. For a given odds ratio it is possible to calculate the joint probabilities of all possible outcomes after the census and CCS. Therefore for any individual it is possible to complete the following 2x2 table of probabilities:

	Counted by CCS	Missed by CCS	
Counted by census	p_{11}	p_{10}	p_{1+}
Missed by census	p_{01}	p_{00}	p_{0+}
	p_{+1}	p_{+0}	1

The values for overall census coverage (p_{1+}) vary for each individual but do not vary across simulations. The values for the CCS response rate (p_{+1}) are fixed for each individual within a simulation but vary across simulations from perfect (100%) to 95% and 80%. The odds ratio is varied from 0.1 (people not in census are ten times more likely to be in the CCS than those counted in the census) to one (independence) to 10 (people in census are ten times more likely to be in the CCS than those not counted in the census). This means that as the odds ratio decreases from one to zero the chance of the two counts finding different people increases (p_{11} and p_{00} go down). Conversely as the odds ratio increases from one, p_{11} increases to its maximum value which is the minimum value of p_{1+} and p_{+1} .

The regression estimator was then used with the union count as described at the beginning of Section 4 and the variances were estimated as before. Performance of the proposed CCS design for different levels of dependence was assessed by computing the relative bias and

relative root mean square error (RRMSE), a combination of variance and bias, across all 1000 simulations. The RRMSE is defined as:

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

and is calculated within each age sex group across all 1000 simulations for each scenario. The relative bias is calculated similarly as:

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

for the same groups. The results are presented in a series of graphs for varying odds ratios by sex. Figures 2 to 4 are for males and show the RRMSE.

(Figures 2 - 4 here)

Figures 2 and 3 show that for odds ratios of 0.1 and one the RRMSE remains below 2.5 per cent even when the CCS response rate is 80 per cent. However, as the odds ratio increases above one the same people tend to be missed by the census and the CCS. In this case, as the CCS response rate falls the RRMSE goes up, especially in those groups where the census coverage is also lower, such as males aged 20-29. The message here is that for a high CCS response rate the regression estimator will still do well regardless of dependence. As the CCS response rate falls, dependence with an odds ratio greater than one will lead to the regression estimator failing. At this stage it is unclear what level of dependence will exist between the census and the CCS. It is likely that it will be greater than one in most areas with the census and CCS tending to miss the same people. However, there is also the argument that those who respond to the census will be less likely to cooperate in the CCS, feeling that their civic duty has been done, than those missed by the census. The results for females have the same general pattern but the variation across age groups is less reflecting the lower levels of female underenumeration at all age groups except the oldest.

Figures 5 to 7 present the relative biases for males for the changing odds ratios. For reference the relative bias for the unadjusted census counts is also presented (termed 'No CCS' in Figures 5 to 7). Figures 5 to 7 show that as the RRMSE increases with the odds ratio the negative bias of the regression estimator also increases. Relative to the unadjusted counts the regression estimator still does very well for odds ratios of 0.1 and one. Comparing Figures 4 and 7 it can be seen that for an odds ratio of 10, once the CCS response rate has fallen to 80 per cent, the RRMSE is almost entirely determined by the bias. This will have serious consequences for the calculation of confidence intervals from estimated variances as these are centred on the value of the estimate and assume that the estimator is unbiased. Therefore, the confidence interval will be calculated around the wrong point. Note, however, that even in this worst case the adjustment procedure is still doing better than not adjusting at all. As before, the results for females show the same pattern but with less variability across the age groups.

(Figures 5 - 7 here)

From this initial sensitivity analysis it can be seen that determining the possible extent and direction of dependency between the census and CCS is important. The most concern is when the odds ratio characterising this dependency is greater than one. As the odds ratio decreases to zero the regression estimator will not suffer, even if the CCS response rate falls, as it will still find the different people for the union count. To see if the union count regression estimator could be improved upon the simulations for a CCS response rate of 80 per cent were re-run. For each sampled postcode three 'Y' counts were calculated:

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

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

Y_{DSE} = DSE applied to the census and CCS counts a postcode.

When the CCS has a perfect response rate all three counts are identical. For correlated non-response in the CCS there is the condition that $Y_{MAX} \leq Y_{UNION} \leq Y_{DSE}$. Population totals were calculated using the regression estimator based on each count. The simulations were run for the same range of odds ratios and from these RRMSE and relative bias were calculated as before.

(Figure 8 here)

Figure 8 shows the full set of graphs for the RRMSE and relative bias for males. Figure 8 demonstrates that the regression estimator based on Y_{MAX} always performs poorly in terms of both RRMSE and relative bias. The regression estimator based on Y_{DSE} performs as expected. For an odds ratio of 0.1 it has a positive relative bias which feeds into the RRMSE making it less efficient than the regression estimator based on Y_{UNION} . For an odds ratio of one (independence between the census and CCS counts) Y_{DSE} performs the best with almost zero bias and good RRMSE. This should be the case as the DSE is based on an independence model. For an odds ratio of 10 Y_{DSE} performs slightly better than Y_{UNION} as one would expect given that $Y_{UNION} \leq Y_{DSE}$ but in all cases the negative bias for young men is quite noticeable.

If one can be sure that the odds ratio is going to be greater than one, that is the census and CCS tend to find the same people, the regression estimator based on Y_{DSE} offers the best protection against negative bias in estimates of the population totals. However, in the situation where well motivated CCS fieldworkers find the missed people, Y_{DSE} will give a clear over estimate. Given that currently one can argue for both situations it is useful to look at other properties of the estimators. Figure 9 presents graphs to show coverage of the ultimate cluster variance estimator for each of the estimators for the range of odds ratios.

(Figure 9 here)

Figure 9 clearly shows the effect the bias has on the coverage given by the variance estimator for Y_{MAX} with it being 50 per cent or less instead of around 95 per cent. For an odds ratio of 0.1 there is not much difference between the other two. However, Y_{DSE} remains slightly conservative for an odds ratio of one while Y_{UNION} slips just below the 95 per cent line. For an odds ratio of 10 the negative bias in all three cases leads to variance estimators not giving correct coverage. The problem is least severe for Y_{DSE} but for young males it is still only giving 25 per cent coverage. Poor coverage in all cases for men aged 85+ year old men reflects the fact that for this age group the variance is often estimated as

zero from the sample. This is due to the fact that as the true counts are small the CCS often finds no additional people. This generates a perfect regression line with no residual variance.

4.2 TSE simulation results

As an initial investigation of using a TSE the standard DSE was extended by assuming independence between all three lists to give the following estimator (see Annex A for the derivation):

$$\text{Total} = \sqrt{\frac{\text{Census} \times \text{CCS} \times \text{Admin.}}{\text{Number in all}}} \quad (7)$$

The TSE approach could not be implemented in the CCS simulations as there is not a suitable administrative list available to match to 1991 Census data. Therefore, to investigate the effect of dependence and non-response when using the TSE a population of 1000 people was simulated and each person was given a probability of being in the census, a probability of being in the CCS and a probability of being on a third administrative list. The odds ratio was used as before to set a level of dependence between the census and CCS. The probability of being on the third list from other administrative sources remained independent of the census and CCS. This situation is considered to be realistic when the administrative list is collected completely independently of the census process and using a different methodology. A multinomial trial was then used to generate the outcome of each individual after the census and CCS (ie counted in the census, not counted in the CCS and found on the administrative list; counted in the CCS, not in the census and not found on the administrative list). From the three lists, the TSE given in (7) was used to estimate the total population. This was repeated 1000 times. The whole simulation was repeated for a variety of odds ratios, with different response rates for the census and CCS.

The final stage introduced list-inflation on the administrative list. This represents people who have been registered twice, died or moved out of the area but have not been removed from the list. This was achieved very simply, by inflating the number of people who only appear on the list, such that the total population given by the list was 102% of the true population (of which 90% were in the true population). The results from the simulations are summarised in Figures 10 to 13 below. They demonstrate the effects of dependence when census coverage equals 90% and 70% respectively with and without list-inflation on the administrative list. A CCS coverage of 85% is expected as a minimum, but may be lower for some types of people. Therefore a range of CCS coverages was used. This response rate is similar to those obtained by the Labour Force Survey (LFS) between 1990 and 1996, which ranged between 77 to 85 per cent (OPCS, 1992, 1996).

(Figures 10-13 here)

Figures 10 and 11 show that for high census coverage the TSE estimates are close to the true population total, even when there is dependence between the census and CCS and a poor CCS response rate. However, the effects of dependence and poor CCS coverage are exaggerated by the lower census coverage of 70%. In this simple simulation variances were not calculated for each population estimate. The error bars given are $\pm 2\sigma$ calculated from the 1000 estimates given by the simulation. These show that for a census coverage of

90% the population estimate is usually less than five per cent away from the truth for a poor CCS response rate and within one per cent for an excellent CCS response rate. In the situation of poor census coverage only a good CCS response rate will get the population estimate within five per cent of the truth unless the census and CCS are basically independent with an odds ratio close to one.

Figures 12 and 13 illustrate the effect of the presence of list inflation on the third independent list. When compared with Figures 10 and 11 it can be seen that list inflation has the effect of inducing a positive bias on the estimation of the total population. It effectively leads to an upward 'shift' in the estimate. Therefore Figures 12 and 13 are similar in pattern to Figures 10 and 11 but moved up the y-axis. The list inflation introduced into the simulations results in an increase in the over-estimation of the total population of approximately 200% for the case of the 90% census coverage, 55% CCS and for odds ratio = 0.2 (1030 increases to 1098). The effect is reduced for the case when the census coverage is 70% (1112 increases to 1179). For the situation when negative bias exists, that is when the odds ratio is equal to five say, the effect of the positive bias due to the list inflation on the third list results in a low net error in the estimate of the total population. However, the gross error in this case is great and this cannot be overlooked.

4.4 Conclusions

The initial conclusions from the simulation study in Section 3.6 are that for a perfect CCS the regression estimator works well, recovering any loss of efficiency due to the two stage design and multivariate stratification. However, the spread of RSEs across the simulations is still quite high. The coverage performance of the confidence intervals based on the ultimate cluster variance estimator is excellent, even though there is a slight positive bias in the regression estimator due to postcodes with zero counts. In addition the regression estimator still performs well in the presence of correlated non-response. As the response rate for the CCS decreases the direction and the extent of the dependence becomes important, especially for those age-sex groups with the lowest census response rates. This shows how vital it is to get a high CCS response, since once this is achieved dependence between the CCS and census counts becomes a side issue. High response is also important for variance estimation since for increasing odds ratios and low response the bias dominates, and the variance of the regression estimator based on either the DSE or union count tends to zero as the CCS finds fewer and fewer of the missed people.

The TSE simulations show that the TSE performs best when coverage for both the census and CCS are high and high coverage reduces the effect of dependence. At low levels of coverage, dependence between the census and the CCS biases the estimate quite severely. The addition of an independent third list may improve the accuracy of the estimate, although this has not been tested here, but finding a suitable source is difficult since the problems of list inflation appear to outweigh any benefits that may be derived from the use of a third list. The magnitude of the actual list inflation present in the administrative sources needs to be investigated.

5. ADJUSTING FOR UNDERENUMERATION WITHIN SMALL CENSUS AREAS

The ultimate aim of the ONC project is a single individual level census database fully adjusted for underenumeration. This requires a procedure that estimates weights at an

individual level to adjust for underenumeration. These weights can then either be applied directly to produce a weighted database or used in the imputation of missing people at a very small area for both counted households and missed households. This section proposes a modelling approach to obtain these weights and briefly discusses the options regarding weighting and imputation.

5.1 Status of Individuals after the Census and CCS

Let us assume that the CCS has taken place in a sample of postcodes within each design level group. Without loss of generality only one design group is considered. For those postcodes in the sample there are two lists of individuals, one from the census and one from the CCS. These lists can, in principle, be matched to produce a single list of individuals containing all those individuals found in the census with any extras from the CCS. This is a slightly different assumption to the one in Section 4 and recognises that the CCS will not find all the people that the census does. The assumption is that no one is missed by both. This is a particularly strong assumption for some areas and work will be needed to assess the robustness of the approach with respect to this assumption.

At the individual level each person has:

- | | | |
|------|--|----------|
| i) | a vector of their socio-economic characteristics
(age, sex, marital status, ethnicity, economic status) | <u>X</u> |
| ii) | a vector of their household characteristics
(tenure, building type, multiple-occupied, number of residents) | <u>Z</u> |
| iii) | an indicator of their household structure | S |

The household structure variable indicates the type of relationship between individuals within the household such as:

- Single person;
- Couple with no children;
- Nuclear family (couple and children);
- Extended family (couple, children and others);
- Single parent family;
- Household of unrelated members;
- Communal establishment (institution).

Each individual i belongs to a household j within a postcode k within an enumeration district l of district m . The CCS does not contain all districts or postcodes so one needs to use the sampled postcodes to estimate underenumeration in the non-sampled postcodes. From the regression estimation procedures, presented in Sections 3 and 4, using the CCS with the possibility of administrative lists, there are gold standard age sex totals at the design group level. By gold standard it is meant that these are counts which have been adjusted for underenumeration using an optimal strategy. The aim is to share the ‘extra’ people amongst the enumeration districts.

5.2 Multinomial model for small area adjustments

If it is assumed that no individual is missed by both the census and CCS, then in relation to the census, a person is either counted, missed in a counted household, or missed in a missed household. This can be represented by the variable Y_{ijklm} for location klm where:

$Y_{ijklm} = 0$ when individual i of household j is counted in the census

$Y_{ijklm} = 1$ when individual i is missed in the census but his/her household j is counted by the census

$Y_{ijklm} = 2$ when individual i and his/her household j are both missed in the census

Outcomes 1 and 2 are only observable in the CCS areas by matching with the CCS. Let

$$\begin{aligned} \Pr(Y_{ijklm} = 0) &= \pi_{0ijklm} = \Pr(i \text{ is counted}) \\ \Pr(Y_{ijklm} = 1) &= \pi_{1ijklm} = \Pr(i \text{ is missed} \cap j \text{ is counted}) \\ \Pr(Y_{ijklm} = 2) &= \pi_{2ijklm} = \Pr(i \text{ is missed} \cap j \text{ is missed}) \\ \pi_{0ijklm} + \pi_{1ijklm} + \pi_{2ijklm} &= 1 \end{aligned} \quad (8)$$

In general these probabilities will depend on the characteristics of the person, household, postcode, etc. Putting aside measurement error problems⁶ the following multilevel multinomial model can then be fitted for the CCS sample postcodes:

$$\begin{aligned} \ln\left(\frac{\pi_{1ijklm}}{\pi_{0ijklm}}\right) &= \alpha_1 + \underline{\beta}_1' \underline{X}_{1ijklm} + \underline{\gamma}_1' \underline{Z}_{1ijklm} + \eta_1 S_{1ijklm} + \lambda_{1lm} + v_{1klm} + \varepsilon_{1ijklm} \\ \ln\left(\frac{\pi_{2ijklm}}{\pi_{0ijklm}}\right) &= \alpha_2 + \underline{\beta}_2' \underline{X}_{2ijklm} + \underline{\gamma}_2' \underline{Z}_{2ijklm} + \eta_2 S_{2ijklm} + \lambda_{2lm} + v_{2klm} + \varepsilon_{2ijklm} \end{aligned} \quad (9)$$

where α_r , $\underline{\beta}_r$, $\underline{\gamma}_r$ and η_r ($r = 1,2$) are the unknown fixed parameters associated with the variables \underline{X} , \underline{Z} and \underline{S} and λ_{rlm} , v_{rklm} and ε_{rijklm} are independent random error terms with $\lambda_{rlm} \sim N(0, \sigma_{rlm}^2)$, $v_{rklm} \sim N(0, \sigma_{rklm}^2)$ and ε_{rijklm} constraining the distribution at the individual level to be multinomial. The model specified by (9) corresponds to a standard random intercepts model and computer packages exist to estimate the unknown fixed parameters as well as the parameters $(\sigma_{rlm}^2, \sigma_{rklm}^2)$ for the distributions of the random effects at the postcode level and enumeration district level. The addition of random effects allow for unexplained heterogeneity between postcodes and enumeration districts in small area estimation, due to unobserved covariates. In general the model specified by (9) can be further extended to also test for significant random coefficients.

This is a similar approach to that taken by the US Census Bureau in 1990 to solve a slightly different problem. In their case the problem was to impute whether an individual found in their Post Enumeration Survey (PES) was counted in the census or not where matching was not conclusive. They used the predicted probability from a hierarchical logistic regression model as the imputed value (rather than 0 or 1) when adding up the number of people counted in the 1990 Census to calculate the undercount adjustments. They didn't use the predicted probabilities as weights to adjust the census. This is detailed in Belin *et al.* (1993) along with the estimation procedures used to fit the regression model.

³ When there is a choice between a census and CCS measure the CCS measure will be used in the modelling.

5.3 Prediction for non-sampled postcodes

Once the model has been fitted, the first step is to use the estimated fixed parameters $\hat{\alpha}_r, \hat{\beta}_r, \hat{\gamma}_r$ and $\hat{\eta}_r$ and estimated high level residuals $\hat{\lambda}_{rlm}$ and \hat{v}_{rklm} to get predicted probabilities for each of the different types of individuals and households in all sampled areas given by:

$$\hat{\pi}_{rijklm} = \frac{\exp(\hat{\alpha}_r + \hat{\beta}_r' \underline{X}_{rijklm} + \hat{\gamma}_r' \underline{Z}_{rjklm} + \hat{\eta}_r S_{rjklm} + \hat{\lambda}_{rlm} + \hat{v}_{rklm})}{1 + \sum_{r=1}^2 \exp(\hat{\alpha}_r + \hat{\beta}_r' \underline{X}_{rijklm} + \hat{\gamma}_r' \underline{Z}_{rjklm} + \hat{\eta}_r S_{rjklm} + \hat{\lambda}_{rlm} + \hat{v}_{rklm})} \quad (10)$$

when $r = 1, 2$ with $\hat{\pi}_{0ijklm} = 1 - \hat{\pi}_{1ijklm} - \hat{\pi}_{2ijklm}$. Extending this to obtain predicted probabilities for individuals in non-sampled postcodes is straightforward for the fixed effects model. In this situation the assumption is made that the model fitted for sampled postcodes holds in all postcodes. For the multilevel model a similar assumption can be made. However, due to the independence assumption made in estimating the parameters of the multilevel model, the estimate of the postcode and enumeration district random effects for the non-sampled postcodes is zero. Intuitively, it would be expected that if these postcodes were in the sample this would not be the case. A solution to this would be to fit a spatial⁷ random effects model. This would lead to non-zero predictions of random effects in non-sampled postcodes. Computationally speaking this is currently extremely difficult. The proposal which is being considered is to fit the model assuming independence for the random effects, and estimate random effects for non-sampled postcodes by averaging the corresponding random effects from the h ‘closest’ sampled locations based on a possibly non-spatial measure of distance. This means that in principle for all areas it is possible to use (10) to estimate $\hat{\pi}_{0ijklm}, \hat{\pi}_{1ijklm}, \hat{\pi}_{2ijklm}$.

5.4 Adjusting the Census

The next stage is to adjust the census counts. Let N_{ijklm} be the census count of individuals with the set of characteristics given by i from households with characteristics given by j in location klm. (eg. 20-24, employed male, married and renting a house within location klm.)

$$\Pr(\text{People with characteristics ij are counted in location klm}) = \hat{\pi}_{0ijklm}$$

$$\text{implies } \Pr(\text{People with characteristics ij are missed in location klm}) = 1 - \hat{\pi}_{0ijklm}$$

From this the number of people with individual and household characteristics ij who are missed in location klm is given by:

⁴ Spatial does not need to mean geographic. It may be more appropriate to ‘borrow strength’ from other areas based on distance measured in terms of demographic characteristics. This reflects the situation where, especially in cities, rich and poor live in contiguous areas.

$$N_{ijklm} \times \left(\frac{1}{\hat{\pi}_{0ijklm}} - 1 \right) = N_{ijklm} \times \left(\frac{1 - \hat{\pi}_{0ijklm}}{\hat{\pi}_{0ijklm}} \right) = \hat{m}_{ijklm} \quad (11)$$

The problem is now how to allocate these ‘extra’ people to already counted households or completely missed households. Given that an individual is missed the probability that their household was missed or counted is required.

$$\Pr(j \text{ is counted} \mid i \text{ is missed}) = \frac{\Pr(j \text{ is counted} \cap i \text{ is missed})}{\Pr(i \text{ is missed})} = \frac{\hat{\pi}_{1ijklm}}{1 - \hat{\pi}_{0ijklm}} \quad (12)$$

$$\Pr(j \text{ is missed} \mid i \text{ is missed}) = \frac{\Pr(j \text{ is missed} \cap i \text{ is missed})}{\Pr(i \text{ is missed})} = \frac{\hat{\pi}_{2ijklm}}{1 - \hat{\pi}_{0ijklm}}$$

From this the estimated number of missed people from counted households is:

$$N_{ijklm} \times \left(\frac{1 - \hat{\pi}_{0ijklm}}{\hat{\pi}_{0ijklm}} \right) \times \left(\frac{\hat{\pi}_{1ijklm}}{1 - \hat{\pi}_{0ijklm}} \right) = \frac{\hat{\pi}_{1ijklm}}{\hat{\pi}_{0ijklm}} \times N_{ijklm} = \hat{m}_{1ijklm} \quad (13)$$

and the estimated number of missed people from missed households is:

$$N_{ijklm} \times \left(\frac{1 - \hat{\pi}_{0ijklm}}{\hat{\pi}_{0ijklm}} \right) \times \left(\frac{\hat{\pi}_{2ijklm}}{1 - \hat{\pi}_{0ijklm}} \right) = \frac{\hat{\pi}_{2ijklm}}{\hat{\pi}_{0ijklm}} \times N_{ijklm} = \hat{m}_{2ijklm} \quad (14)$$

where $\hat{m}_{ijklm} = \hat{m}_{1ijklm} + \hat{m}_{2ijklm}$.

5.5 Imputing extra people verses weighting

The current proposal is that these missing individuals with characteristics ij in location klm be recreated by imputing synthetic records in the census database for location klm . At present research is underway as to how this imputation procedure will proceed. It is likely that hotdeck methodology will be used which will tie in with the imputation procedures being developed by ONS for imputing missing values in 2001. Of particular concern is the problem of finding donor households for people missed from counted households. In this situation there is the danger of distorting household relationships. An example would be giving a single mother a husband while the married woman whose husband was missed remains single. In this case there are right number of single parents but not necessarily with the correct characteristics. This issue is still being researched.

For the people from the missed households, there will be a set of groups of people given by the different \hat{m}_{2ijklm} . The task is then to fit the individuals back together as households. It is likely that this will require modelling to estimate the number of households missed so that the individuals are formed into the correct number of additional households. Again, this issue still requires further research.

One possible way of avoiding these issues is to create a weighted census database. The weights are obtained directly as the inverse of the predicted probabilities calculated in Section 5.4. The possibility of using weighted tables for census output is also being investigated as an alternative to imputation. This will also require the production of household weights for the census tables at a household rather than individual level.

5.6 Simulation Study Methodology

The same underlying method used for the CCS design simulation in Section 3.6 was used here. Each individual in the true population had the same probability of being counted in the census. Initially 10 censuses, each with its own CCS, were simulated which was fewer than in the CCS design simulation due to the need to keep the individual level data. This is computationally much more time consuming than the totals needed for the county level estimation. For the simulations presented here the CCS was assumed to have perfect coverage. However, in general this will not be the case. It is intended that further simulations will be carried out to investigate the effect of correlated non-response in the CCS. The simulation presented here is used to demonstrate that the proposed model is sensible and warrants further investigation.

For each census-CCS pair, a matching procedure was carried out to determine the response state (8) to which each individual belonged. The fixed effects version of the multinomial model (9) was fitted to each pair. The explanatory variables used were age group, sex, and HtC index. This was the same HtC index as that described in Section 3.

Once the model was fitted for a particular census the predicted probabilities were calculated using (10) simplified for a fixed effects model. $\hat{\pi}_{0ijklm}$ (the probability of being counted) was used to make an adjustment for *all* missed people within each enumeration districts since $1 - \hat{\pi}_{0ijklm}$ is the probability of being missed in the census. This gave enumeration district counts adjusted by age, sex and HtC index for each census. Initial analysis showed very little variation between censuses. Root Mean Squared Error (RMSE) was used to assess the overall performance of the adjustment procedure. This is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^{10} \sum_{i \in d} (\text{observed}_{ij} - \text{truth}_{ij})^2} \quad (15)$$

where j is summed over the ten simulations, i is summed over the enumeration districts within HtC index group d , and n is the total number of enumeration districts in the double sum. In the formula, the observed count can either be the adjusted census count or the unadjusted census count. Similarly, the Bias is calculated as:

$$\text{BIAS} = \frac{1}{n} \sum_{j=1}^{10} \sum_{i \in d} (\text{observed}_{ij} - \text{truth}_{ij}) \quad (16)$$

Again, the observed count can be either the adjusted count or the unadjusted census count. When using the unadjusted census count it should be remembered that the simulation forces each census to have a negative bias due to the fact that people are missed but no

overcount is simulated. However, for the adjusted count it is possible to have zero bias when averaging as for any particular CCS the bias can be positive as well as negative.

5.7 Discussion of Results

The following discussion is preliminary and only present results for the overall adjustment described in Section 5.6, not adjustments split by counted and missed households. Its role is to demonstrate that the concept works, not to give a definitive picture of how the procedure would work in a One Number Census. The above measures have been calculated from the 10 fitted versions of the multinomial model (5) resulting from the simulation. Tables 7 and 8 present the results for enumeration district totals computed at each step of the simulation.

(Tables 7 and 8 here)

Table 7 shows that adjusting the counts reduces the RMSE in each of the HtC categories. It also reduces the difference across the HtC categories. Table 8 shows a dramatic change with the adjustment reducing the bias to very close to zero. Again the difference across HtC categories has also been reduced. Comparing the two tables it can be seen that for the census the RMSE is nearly all due to bias whereas for the adjusted counts it is nearly all variance. This is encouraging as variance is much easier to estimate than bias.

Obtaining the correct overall totals is important. However, adjustments are required by age and sex for a One Number Census as these are known to be key variables by which underenumeration varies. Figures 14 and 15 show the results for the hardest to count group. The adjusted figures are compared to the unadjusted figures so that the gain from adjustment can be seen.

(Figures 14 and 15 here)

Figure 14 shows that in terms of RMSE, the adjustment process is never worse than the unadjusted counts and usually better. For young males Figure 14 clearly shows the added value of the adjustments. The only exception is the 85+ males where the RMSE for the census drops just below the adjusted counts. Figure 15 investigates the bias and it can be seen that for this age-sex group the census approaches zero while the adjustment imputes too many people. In general, in terms of bias the adjusted counts are also better.

6. DISCUSSION AND CONCLUSIONS

This paper has described the research so far undertaken to develop a strategy for a 'One Number Census' in the United Kingdom in 2001. The major tool of this strategy will be a post enumeration survey, to be called the Census Coverage Survey, which will constitute a radical change from previous censuses. It will focus purely on coverage and will comprise a short questionnaire addressing the characteristics most associated with underenumeration. The survey will aim to make estimates, by age and sex, of underenumeration for around 40 subnational areas. The paper has described the design of this survey and estimated sample sizes for a number of levels of precision. The simulation studies in Section 3.6 and 4.1 show that for reasonable CCS response rates accurate subnational population estimates, adjusted for underenumeration can be obtained.

Given agreed subnational estimates of underenumeration it will then be necessary to make estimates at an individual level. The paper describes a multilevel regression approach to estimate an individuals probability of being counted in the census. The results of a simulation study to assess the potential this regression approach are very promising. They show that in the simple case it is possible adjust for missing people at the enumeration district level. The next stage is to introduce more variables and random effects into the modelling. Applying these probabilities to fully adjust the census database through imputation or weighting is an area that still requires extensive research.

The strategy described in this paper will be subject to consultation with the census user community. A final agreed strategy will then be tested in the 1999 Census Dress Rehearsal. Key to this will be the development of the practical aspects of the fieldwork, data capture and matching of census and coverage survey response.

ANNEX A. DERIVATION OF THE TRIPLE SYSTEM ESTIMATOR

Assume three independent random variables C, S, A such that:

$$\Pr(C = 1) = \text{Probability of being counted in the census} = \frac{\text{census count}}{\text{population total}} = \frac{X_{1++}}{X_{+++}}$$

$$\Pr(S = 1) = \text{Probability of being counted in the CCS} = \frac{\text{CCS count}}{\text{population total}} = \frac{X_{+1+}}{X_{+++}}$$

$$\Pr(A = 1) = \text{Probability of being on the administrative list} = \frac{\text{admin. count}}{\text{population total}} = \frac{X_{++1}}{X_{+++}}$$

Under independence:

$$\Pr(C = 1 \cap S = 1 \cap A = 1) = \Pr(C = 1) \times \Pr(S = 1) \times \Pr(A = 1)$$

The above can be written as:

$$\frac{X_{111}}{X_{+++}} = \frac{X_{1++}}{X_{+++}} \times \frac{X_{+1+}}{X_{+++}} \times \frac{X_{++1}}{X_{+++}}$$

Rearranging this equation gives an estimator for the unknown population total of:

$$X_{+++} = \sqrt{\frac{X_{1++} X_{+1+} X_{++1}}{X_{111}}}$$

$$\text{i.e. Total} = \sqrt{\frac{\text{Census} \times \text{CCS} \times \text{Admin.}}{\text{Number in all}}}$$

7. REFERENCES

- Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L. and Zaslavsky, A. M. (1993) Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *J.A.S.A.*, **88**, 1149-1159.
- Charlton, J., Chappell, R. and Diamond, I. D. (1997) Demographic analyses in support of a One Number Census. *Proceedings of the Statistics Canada Symposium*.
- Darroch, J. N., Fienberg, S. E., Glonek, F. V. and Junker, B. W. (1993) A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J.A.S.A.*, **88**, 1137-1148.

- Ericksen, E. P. (1973) A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, **10**, 137-160.
- Ericksen, E. P. (1974) A regression method for estimating population changes of local areas. *J.A.S.A.*, **69**, 867-875.
- Heady, P., Smith, S. and Avery, V. (1994) 1991 *Census Validation Survey: coverage report*, London:HMSO.
- Hogan, H. (1993) The 1990 post-enumeration survey:operations and results. *J.A.S.A.*, **88**, 1047-1060.
- Kendrick, S. (1997) The development of record linkage in Scotland: the responsive application of probability matching. *Proceedings of the 1997 Record Linkage Workshop*, Washington D.C., March 20-21st 1997.
- OPCS (1992) Labour Force Survey 1990 and 1991, Series LFS no 9, London:HMSO.
- OPCS (1993) Rebasng the annual population estimates. *Population Trends*, **73**, 27-31.
- OPCS (1994) Undercoverage in Great Britain. *1991 Census User Guide 58*, London: HMSO.
- ONS (1996) Labour Force Survey User Guide Volume 1, London:HMSO.
- ONS (1998) One Number Census Consultation Paper. Submitted for publication.
- Royall, R. M. (1970) On finite population sampling under certain linear regression models. *Biometrika*, **57**, 377-387.
- Royall, R. M. and Cumberland, W. G. (1978) Variance estimation in finite population sampling. *J.A.S.A.*, **73**, 351-361.
- SAS Institute Inc. (1990) The CLUSTER Procedure: Clustering Methods. In *SAS/STAT Users Guide Version 6*, 4th edn, Volume 1 pp. 529-536. Cary, NC: SAS Institute Inc.
- Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *J.A.S.A.*, **77**, 848-854.
- Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends*, **90**, 31-39.
- Zaslavsky, A. M. and Wolfgang, G. S. (1993) Triple-system modelling of census, post-enumeration survey, and administrative-list data. *J. Business & Econom. Stat.*, **11**, 279-288.

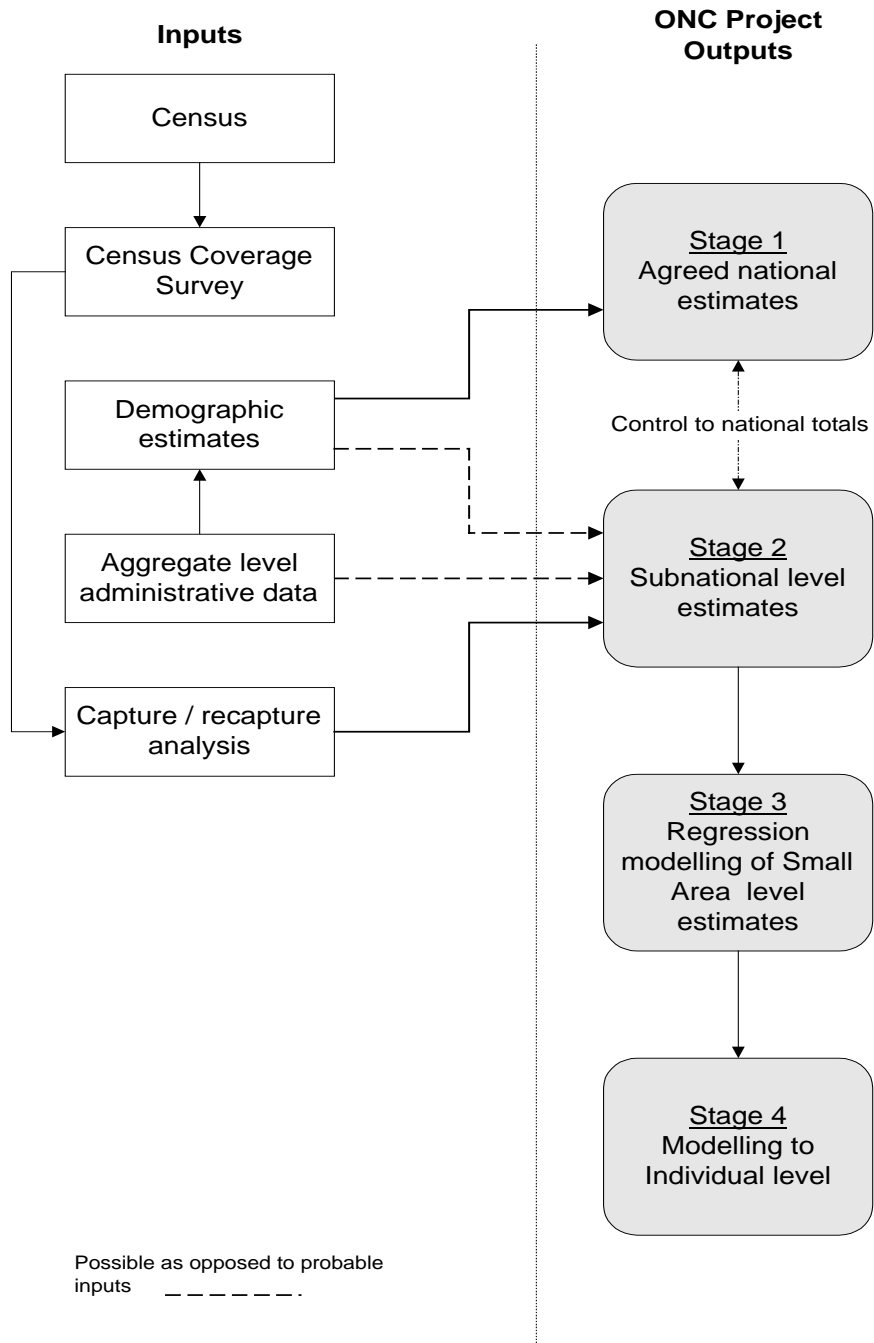


Figure 1. The stages of a One Number Census

TABLE 1
Distribution of 1991 Hampshire enumeration districts by HtC index

Hardness To Count	Number of Enumeration Districts (HtC Index)
Very Easy	249
Easy	626
Medium	874
Hard	925
Very Hard	555

TABLE 2
Sample allocation for the first stage sample in Hampshire

Index Group	Population Size	Number of Size Strata	Sample Size	Outlying ^b
Very Hard	246	15	27	3
Hard	623	35	59	1
Medium	863	35	80	2
Easy	918	35	86	3
Very Easy	551	30	56	2
Outlying ^a	28	-	28	-
TOTAL	3229	150	336	11

a. Enumeration districts classified as outlying due to the size of their male population aged 20-34.
b. Enumeration districts classified in single district clusters by the clustering algorithm.

TABLE 3
Sample allocation for the first stage sample in Kent and West Yorkshire

RSE	Strata	Sample	Outliers	Total Sample
KENT - 3158 EDs				
1.0	190	268	43	311
1.5	122	162	36	198
2.0	105	123	31	154
WEST YORKSHIRE - 4098 EDs				
1.0	125	314	36	350
1.5	100	171	33	204
2.0	100	122	33	155

TABLE 4
Distribution of enumeration districts by HtC index

Hardness To Count	Number of Enumeration Districts
Very Easy	144
Easy	210
Medium	186
Hard	193
Very Hard	197

TABLE 5
Sample allocation for the first stage sample

Index Group	Population Size	Number of Size Strata	Sample Size	Outliers ^b
Very Hard	144	10	12	0
Hard	210	16	17	0
Medium	185	14	14	3
Easy	192	15	18	3
Very Easy	197	15	16	0
Outliers ^a	2	-	2	-
TOTAL	930	70	79	6

a. Enumeration districts classified as outlying due to the size of their male population aged 20-34.

b. Enumeration districts classified in single district clusters by the clustering algorithm.

TABLE 6
Mean Relative Standard Errors for 1000 simulated CCSs

Males				Females			
Age Group	Number of CCSs	Design RSE	Average ^b Estimated RSE	Age Group	Number of CCSs	Design RSE	Average ^b Estimated RSE
0-4	1000	2.73	2.07 (0.593)	0-4	1000	2.66	2.03 (0.695)
5-9	1000	3.86	2.45 (0.754)	5-9	1000	3.92	2.32 (0.745)
10-14	1000	4.52	2.28 (0.734)	10-14	1000	4.45	2.22 (0.687)
15-19	1000	4.45	2.11 (0.645)	15-19	1000	4.19	1.69 (0.589)
20-24	1000	3.33	2.44 (0.613)	20-24	1000	3.22	1.62 (0.483)
25-25	1000	3.02	2.33 (0.508)	25-25	1000	2.99	1.58 (0.407)
30-34	1000	2.92	2.06 (0.471)	30-34	1000	3.12	1.69 (0.423)
35-39	1000	3.94	1.86 (0.466)	35-39	1000	4.04	1.56 (0.433)
40-44	1000	4.18	1.49 (0.406)	40-44	1000	4.53	1.23 (0.418)
45-79	1000	2.83	0.48 (0.168)	45-79	1000	2.77	0.36 (0.139)
80-84	904 ^a	7.67	2.18 (0.978)	80-84	997 ^a	6.24	1.67 (0.588)
85+	721 ^a	10.43	3.71 (1.661)	85+	999 ^a	3.33	2.65 (0.843)

a. Calculation of the variance is not always possible due to zero postcode counts in the CCS.

b. The estimated standard deviation for the distribution of the RSE is given in brackets.

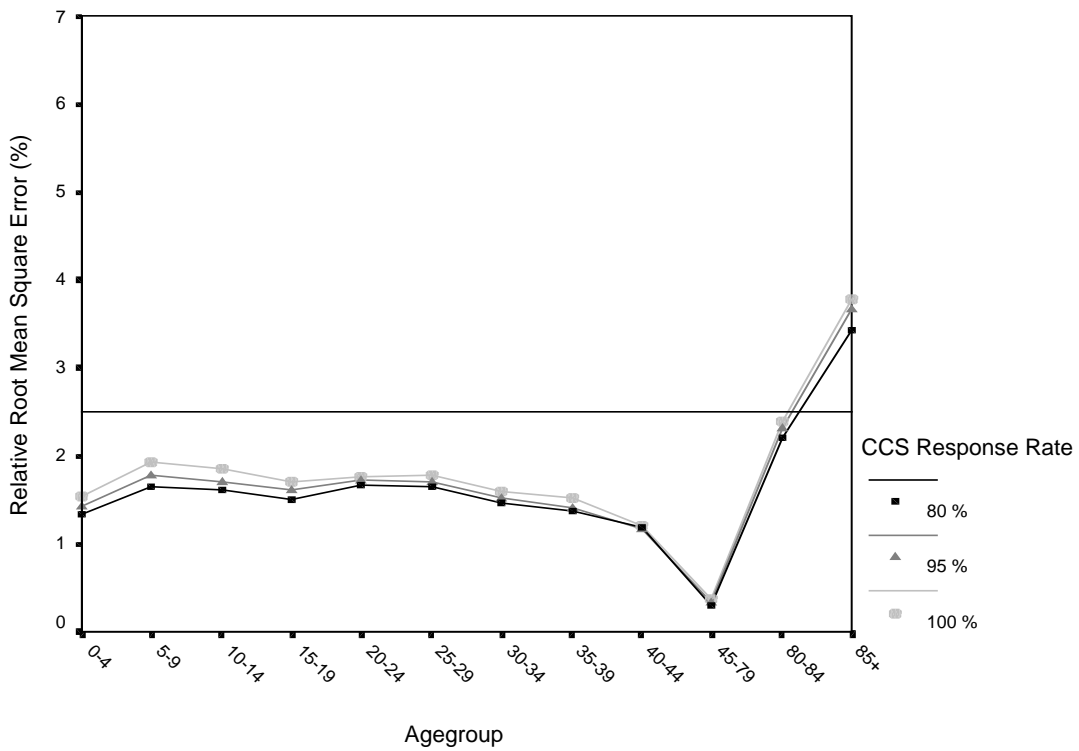


Figure 2. Performance of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 0.1

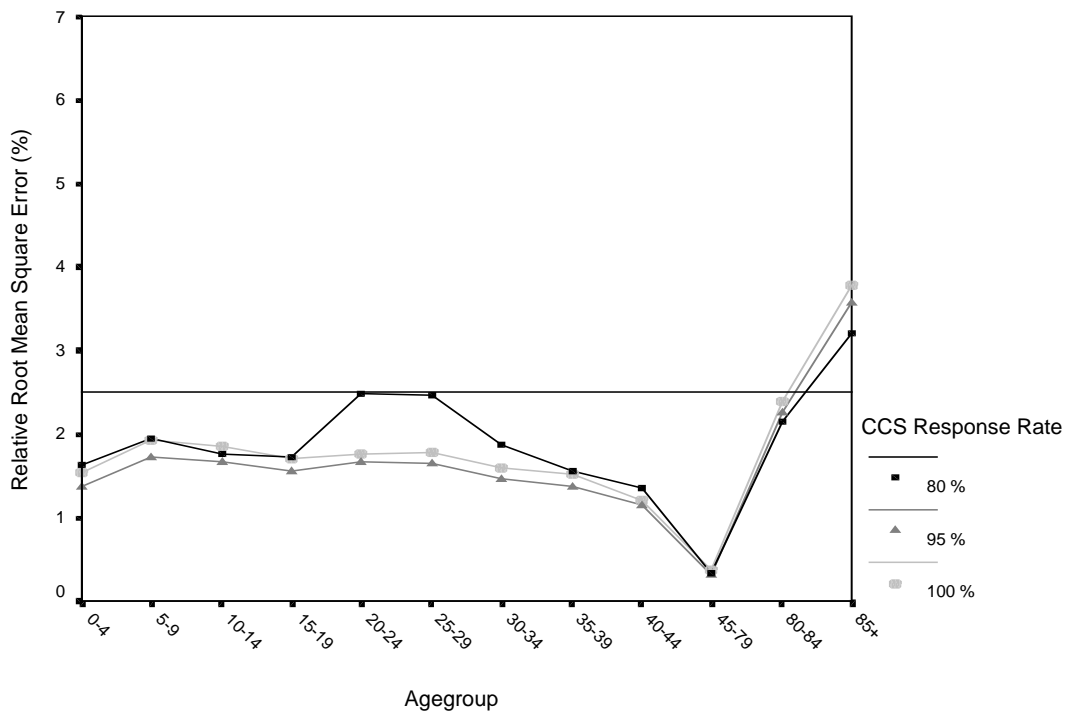


Figure 3. Performance of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 1.0

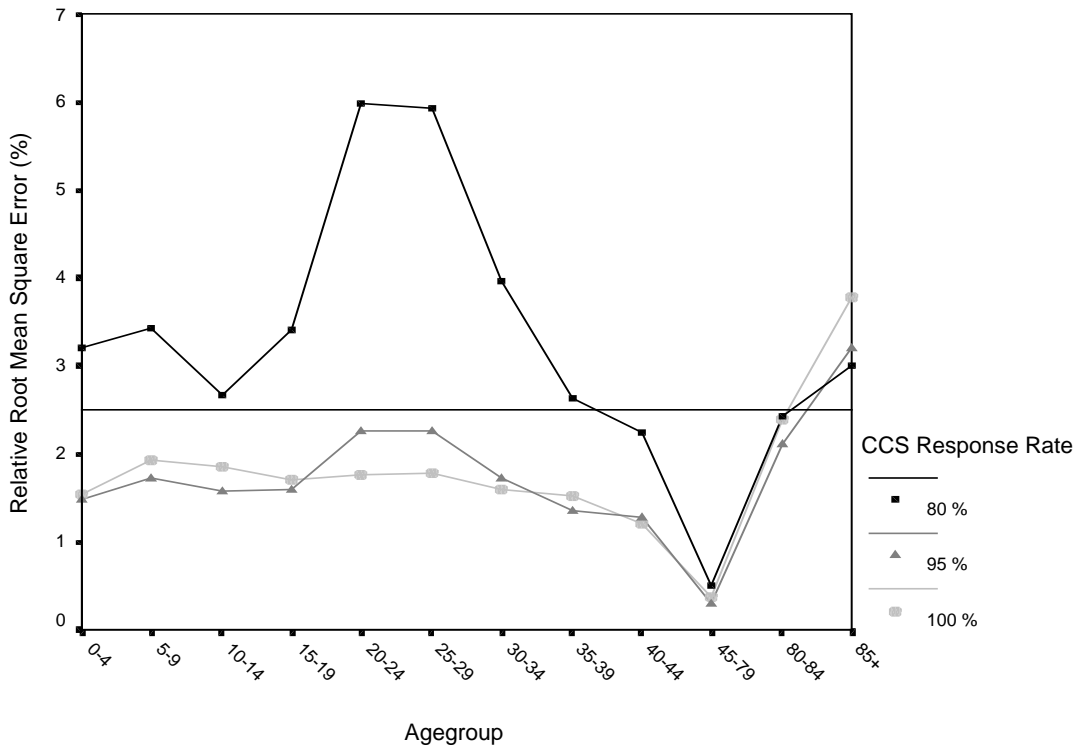


Figure 4. Performance of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 10

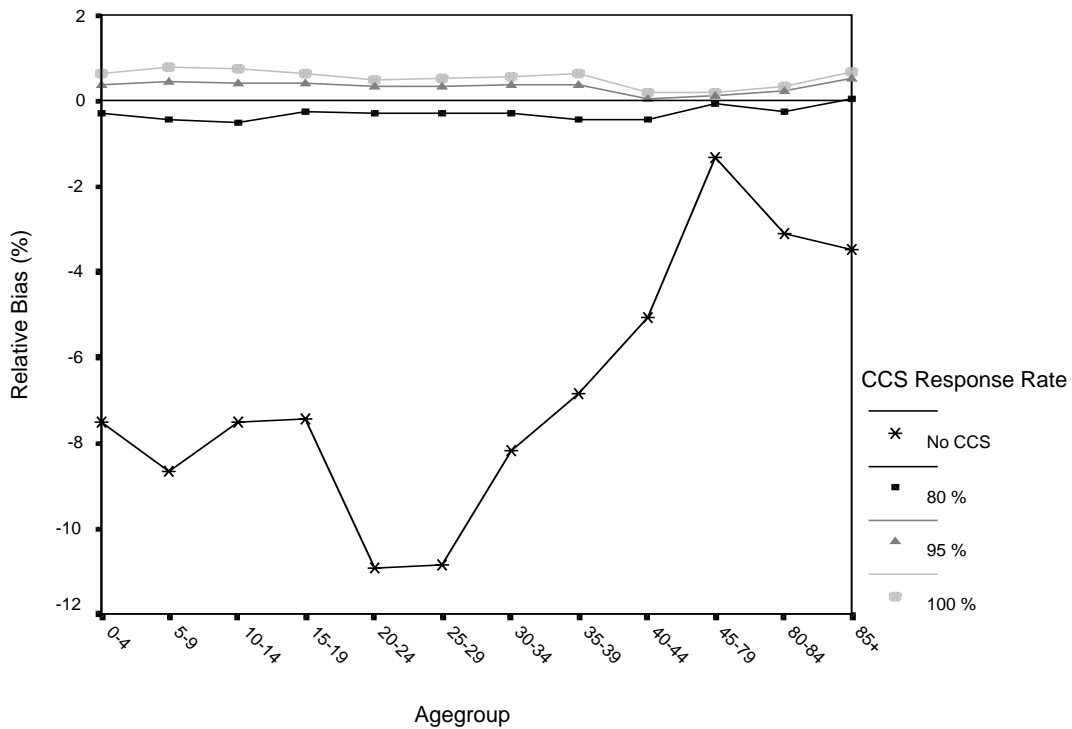


Figure 5. Relative bias of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 0.1

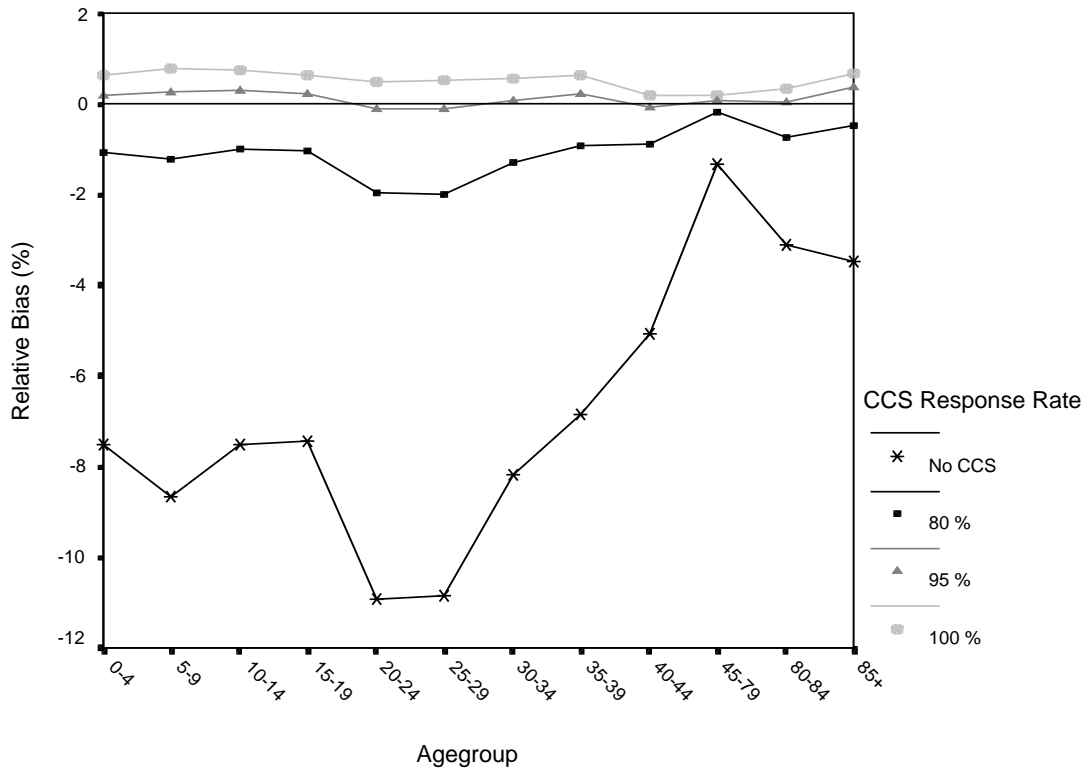


Figure 6. Relative bias of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 1.0

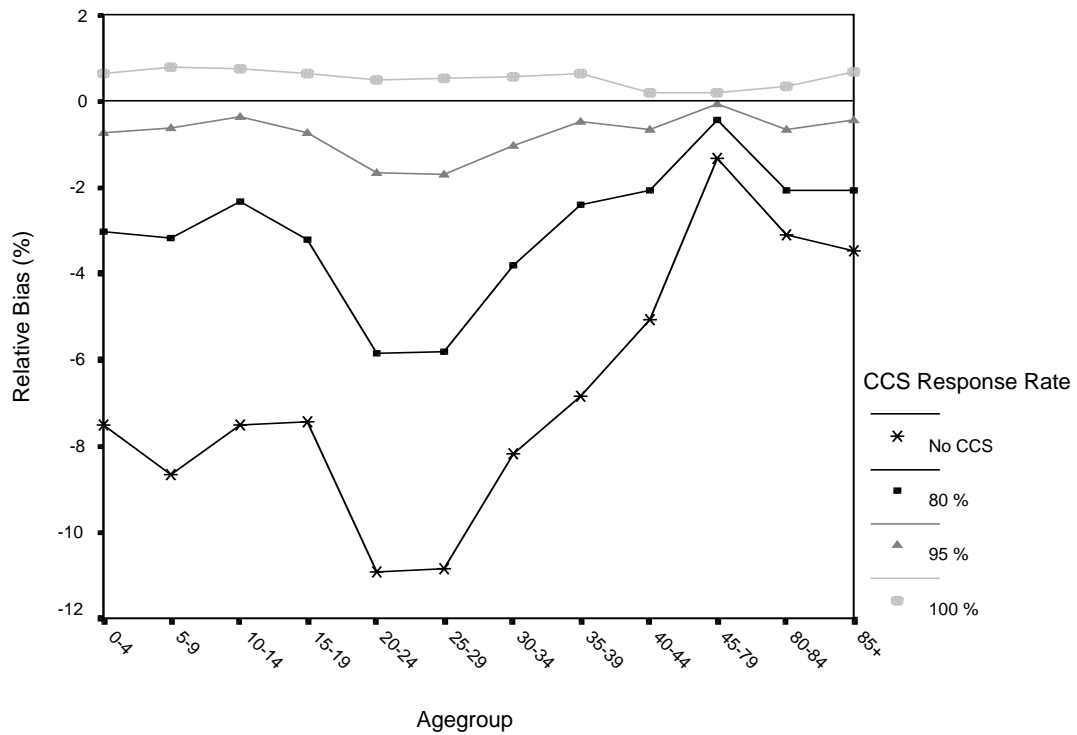


Figure 7. Relative bias of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 10

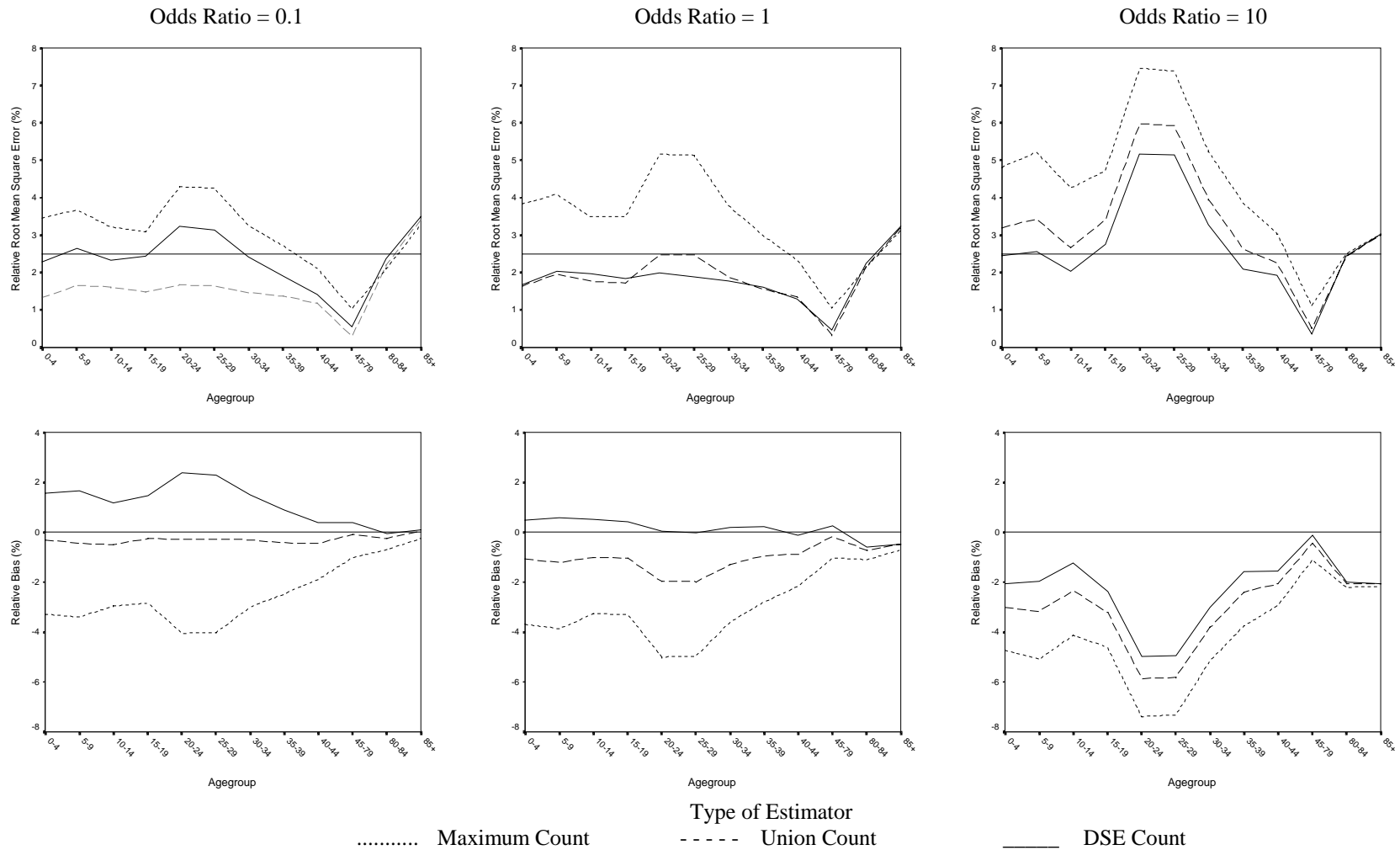


Figure 8. Performance of adjusted county totals for males by type of count used in the regression estimator for varying odds ratios: CCS response rate = 80 %

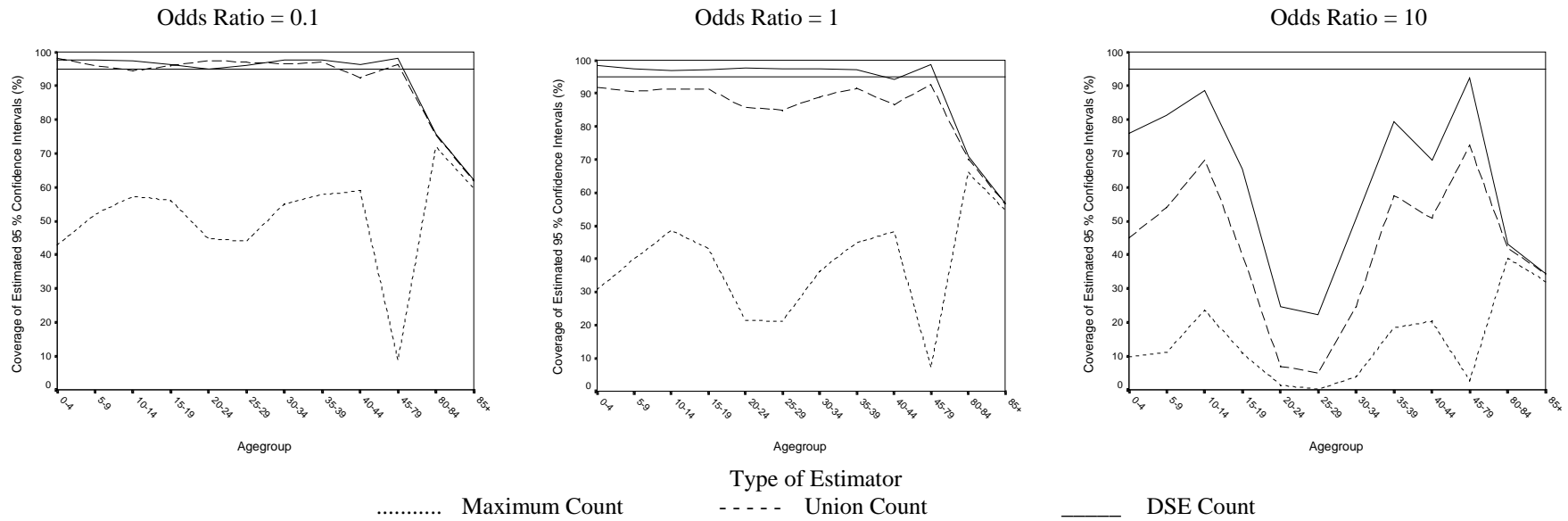


Figure 9. Performance of the adjusted county totals variance estimator for males by type of count used in the regression estimator for varying odds ratios: CCS response rate = 80 %

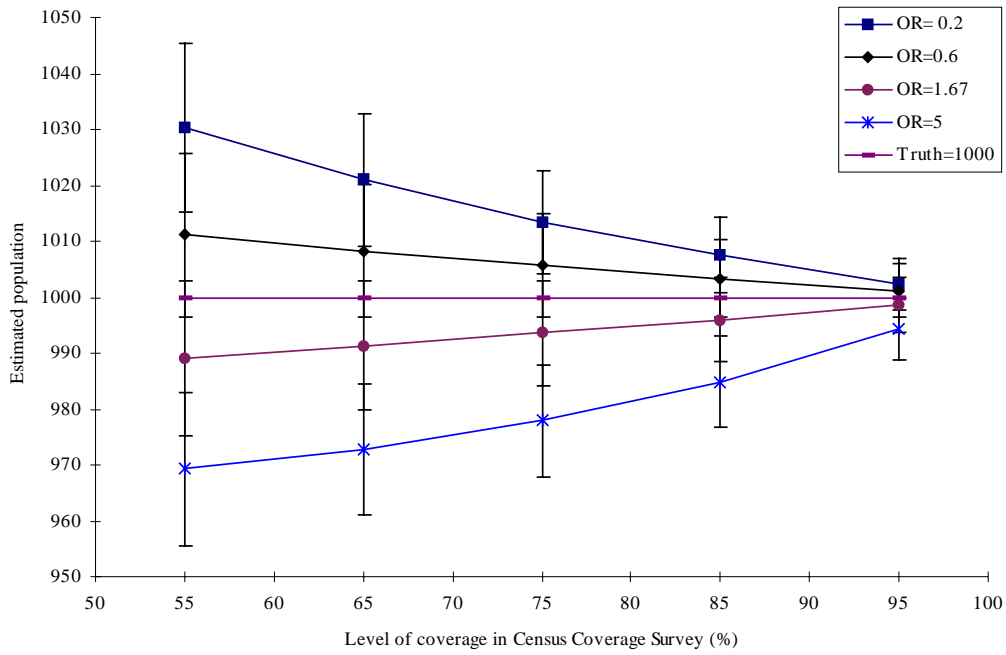


Figure 10. Estimated total population with simulated dependence using a TSE: census coverage = 90%.

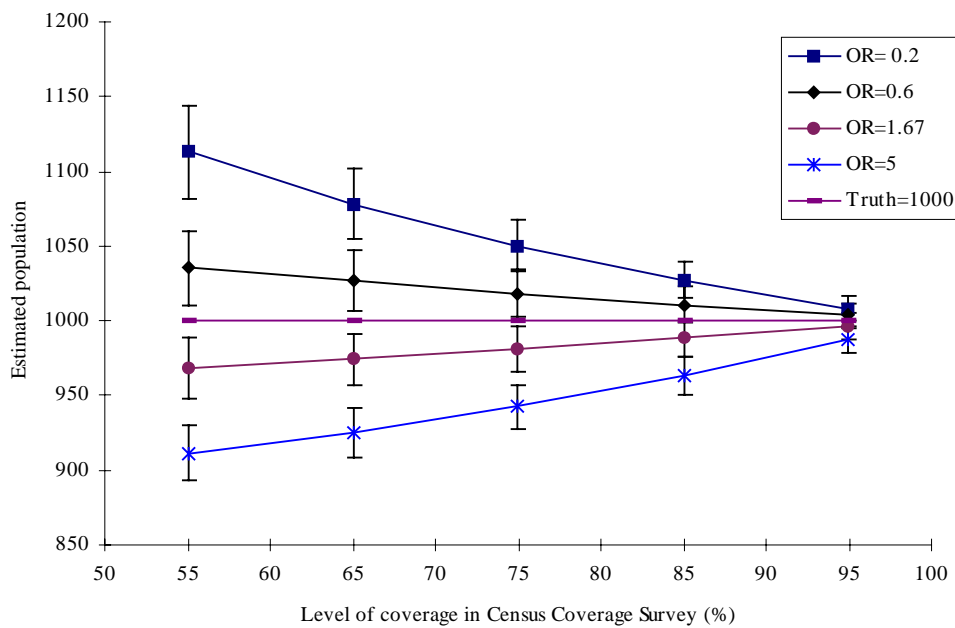


Figure 11. Estimated total population with simulated dependence using a TSE: census coverage = 70%

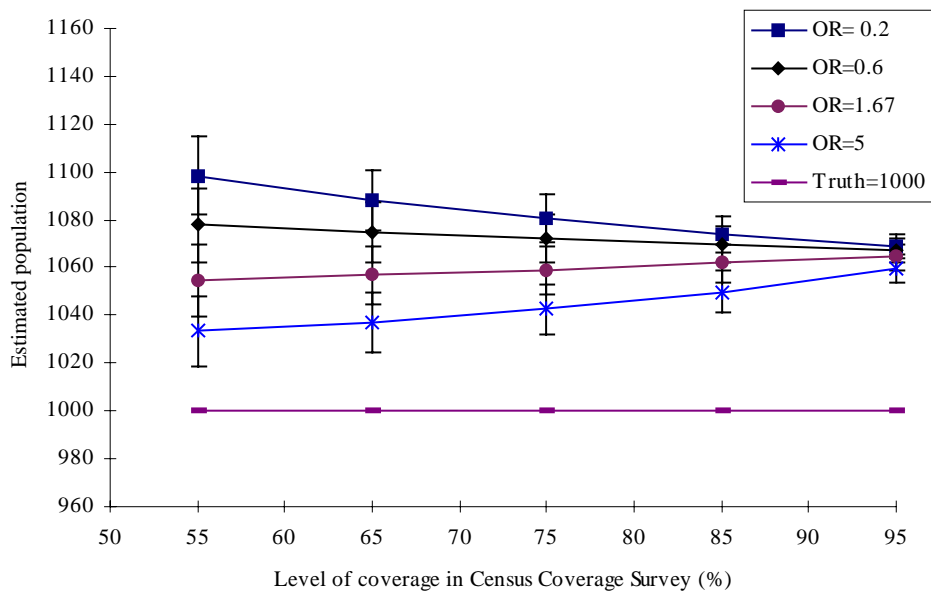


Figure 12. Estimated total population with simulated dependence and list inflation using a TSE: census coverage = 90%

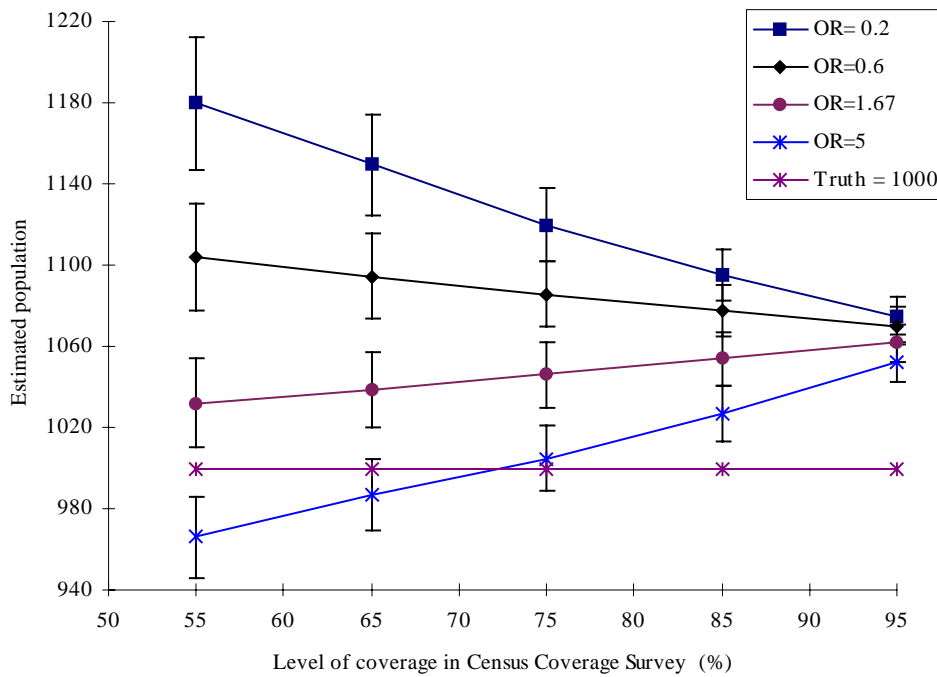


Figure 13. Estimated total population with simulated dependence and list inflation using a TSE: census coverage = 70%

TABLE 7
Root Mean Square Error across all simulations for Enumeration District totals

	HtC Index				
	Very easy	Easy	Medium	Hard	Very hard
Census count	9.94	12.09	13.23	15.92	23.89
Adjusted count	3.64	3.90	4.25	4.42	5.84

TABLE 8
Bias across all simulations for enumeration district totals

	HtC index				
	Very easy	Easy	Medium	Hard	Very hard
Census count	-9.25	-11.22	-12.29	-14.86	-21.84
Adjusted count	-0.08	-0.66	-0.65	-0.01	-0.90

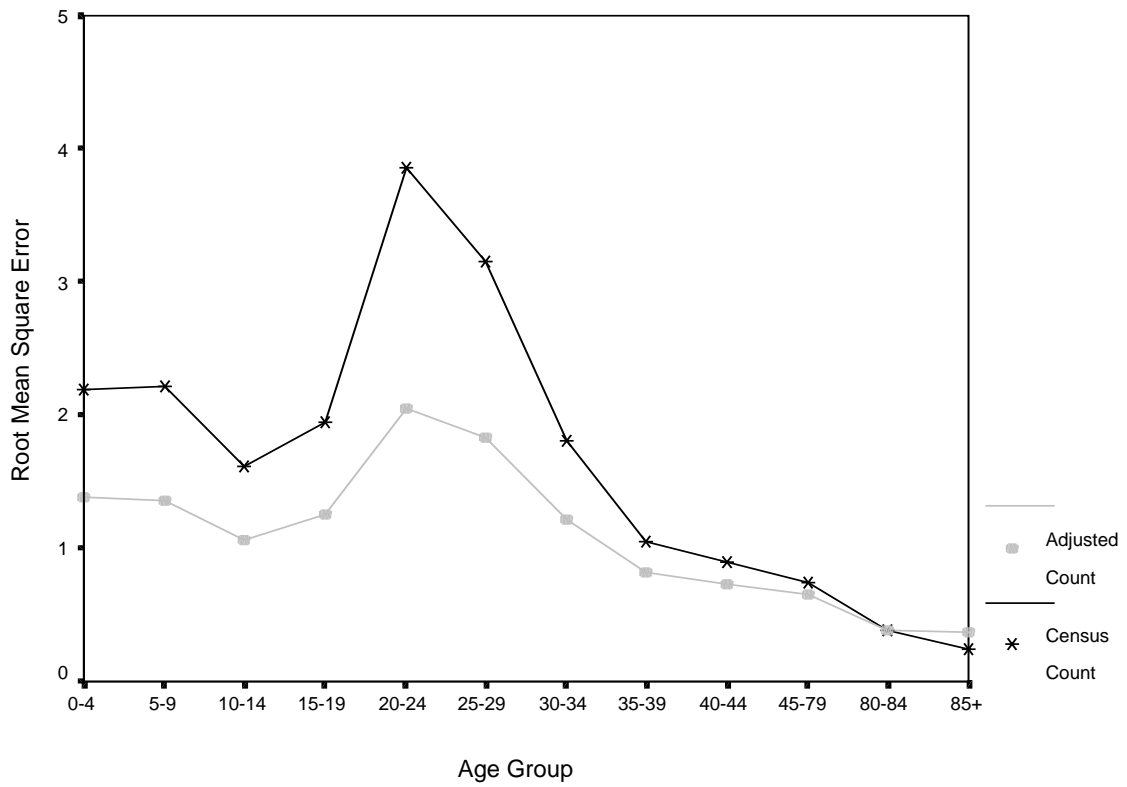


Figure 14. Performance of adjusted enumeration district totals for males relative to unadjusted census totals: HtC = 5 ('very hard')

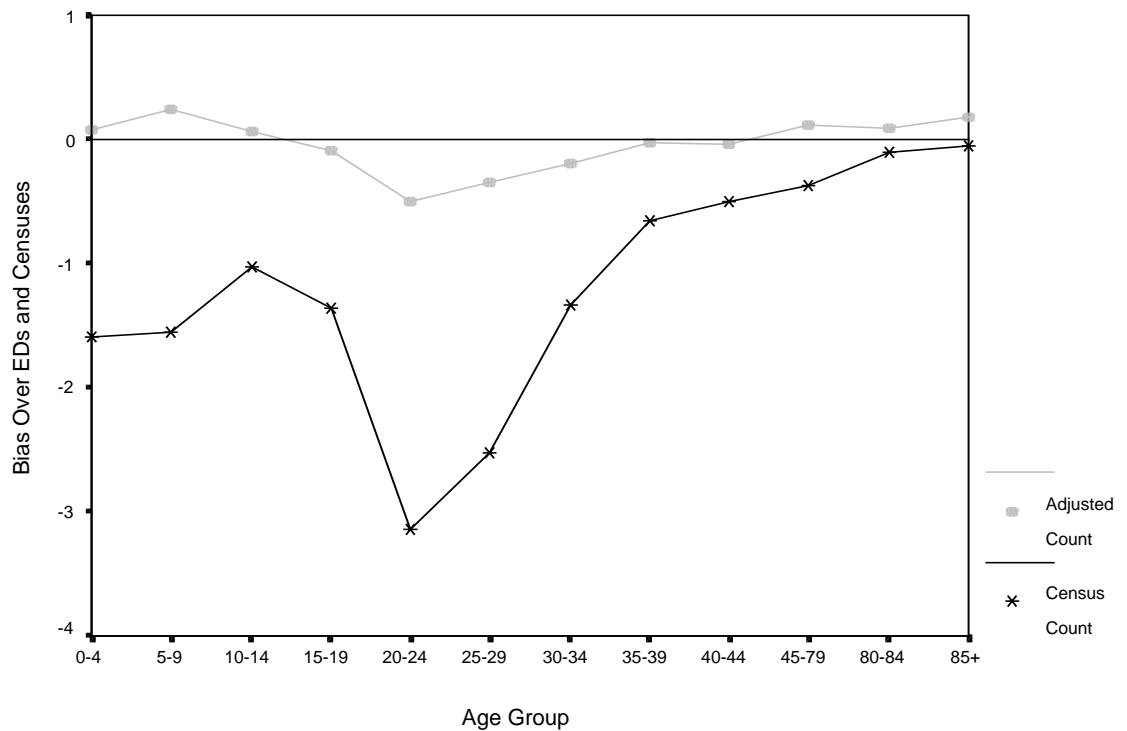


Figure 15. Bias of adjusted enumeration district totals for males relative to unadjusted census totals: HtC = 5 ('very hard')

