

# 2001 - A ONE NUMBER CENSUS

## Contents

	<b>Page</b>
<b>Preface by Dr Tim Holt</b>	<b>3</b>
<b>Chapter 1. Introduction - What is a One Number Census ?</b>	<b>4</b>
Background - Why a ONC in 2001 ?	4
Structure of the paper	5
<b>Chapter 2. Overview of the One Number Census Methodology</b>	<b>5</b>
Maximising coverage in the 2001 Census	5
Overview of the One Number Census Methodology	6
<b>Chapter 3. Management and Consultation arrangements</b>	<b>9</b>
Timing of the ONC process and availability of census statistics	9
Organisation and Steering Committee	10
<b>Annex A. Undercoverage in the 1991 Census</b>	<b>11</b>
<b>Annex B. Design of the Census Coverage Survey</b>	<b>12</b>
<b>Annex C. Dependence, non-response and underenumeration estimation methods</b>	<b>23</b>
<b>Annex D. Demographic analysis in support of a One Number Census</b>	<b>33</b>
<b>Annex E. Modelling down to small areas</b>	<b>43</b>
<b>References</b>	<b>51</b>

## **Preface**

*The purpose of the One Number Census process is to provide Local Authority District level population estimates by age-sex groups that have been adjusted for the undercount in the 2001 Census and to adjust the census database for this undercount at an individual level so that all census outputs add up to 'One Number'.*

I regard this as a very important project. Considerable amounts of resources (approximately £75 billion) are distributed each year from central to local government on the basis of census results and the population estimates. It is vital that the census gives us an accurate picture of society to enable these resources to be targeted effectively. We not only have to ensure that the census itself is as accurate as possible but that any inaccuracies - measured most noticeably by the people not counted - are estimated as accurately as possible.

Those of us in the census taking and user community will remember the problems we encountered with the estimated undercount in the 1991 Census, and particularly the fact that some groups of the population were missed to a greater extent than others. A considerable amount of research has been carried out into not only re-designing the census but also the measurement of the undercount to tackle this problem.

The One Number Census will allow all counts from the census to add to 'One Number' - the national population estimate. This goal presents us with a major methodological challenge. In meeting the aims of the project, it is vital that the proposed methodology is acceptable. As such, I would welcome comments on the proposed methodology as described in this paper.

**Please send your comments, by 1 June 1998 to:**

**Lisa Buckner  
One Number Census  
Census Division  
Office for National Statistics  
Room 4200W  
Segensworth Road  
Titchfield, Fareham  
Hampshire, PO15 5RR**

**Dr D Holt  
Director ONS, Registrar General for England and Wales, and Head of the  
Government Statistical Service**

## **1. Introduction - What is a One Number Census ?**

1.1 The number one objective of the Census Offices<sup>1</sup> for the 2001 Census is to maximise the count of people. The methodology for taking the census is being reassessed with the aim of using resources to their best effect to ensure that the maximum coverage is achieved, and in particular that the differential effect of any undercount is minimised. However, as with all Censuses in the world, there will be some people who are not counted.

1.2 The aim of the One Number Census (ONC) project is to integrate the census and the estimated undercount. Firstly - and essentially - to provide a new base for the mid-year population estimates at the local authority district level, and secondly to adjust the Census database itself for the estimated undercount so that all statistics add to 'One Number' - the national rebased estimate of the population. This has entailed a re-think of the design of the post enumeration survey, now to be known as the *Census Coverage Survey*, concentrating solely on identifying people not counted in the Census. Other indicators of possible undercount, provided by administrative records (at an aggregate level) and demographic analysis, will also be used.

### **Background - Why a ONC in 2001?**

1.3 One of the major uses of the censuses in the UK is in providing a new base for the annual estimates of the population by age and sex. This base needs to take into account the level of underenumeration in the census. This 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.

1.4 In 1991, the level of Census underenumeration at 2.2 per cent (GB - national estimate) was higher than previous censuses, and, in Great Britain, the Census Validation Survey (CVS) - as the post-enumeration survey was known - did not identify the extent and distribution of the undercount. The CVS suggested a total of 290 thousand people were missed by the Census in England, Wales and Scotland, whereas the estimate based on the 1981 Census (adjusted 1981 Census count plus births minus deaths plus net migration in the intercensal period) indicated that 1.2 million people (2.2 per cent) were missed. In addition, underenumeration did not occur uniformly across all socio-demographic groups and parts of the country (for example, the undercount of young males in inner cities was estimated to be greater than 20 per cent) (OPCS 1994).

1.5 It was decided, on balance, that at the national level the 1981-based population estimate was the more reliable national estimate for Great Britain (see OPCS 1993 and 1994 for further details). Several differing population counts became available for 1991, including the unadjusted Census count, the Census counts uprated in line with the CVS, and the 1981 Census based population estimate. This caused difficulties for customers some of whom were obliged to revise their work. It was unclear whether attempts should be made to adjust census statistics and how the differences between the rebased population estimates and the

---

<sup>1</sup> Office for National Statistics (England & Wales), General Register Office (Scotland) and Northern Ireland Statistics and Research Agency

census counts should be interpreted. There was little information on which to base the distribution of the undercount to local areas.

## **Structure of the paper**

1.6 This paper outlines the proposed methodology to counteract the type of problems encountered with the 1991 Census and to achieve the aims stated at 1.2. Chapter 2 provides an overview of the methodology - further detail is given in Annexes A to E, while Chapter 3 describes the consultation process. Note that many of the references to the 1991 Census refer to the Censuses carried out in England and Wales, and in Scotland. The methodology for a ONC in 2001 is, though, being developed for potential application in all four countries of the UK.

## **2. Overview of the One Number Census Methodology**

2.1 This chapter firstly describes the initiatives and changes in methodology being implemented to maximise the coverage of the Census itself, and secondly the proposed methodology to measure and adjust for the undercount.

### **Maximising coverage in the 2001 Census**

2.2 Maximising coverage in the census is seen as the first step towards producing a successful ONC. High coverage in the census is important in itself but also to the accuracy of estimation of the undercount. Details of factors which may have contributed to the undercount in 1991 are presented at Annex A.

2.3 A major Census Test in 1997 provided an opportunity to try out new methods and ideas such as:

- the use of postal methods of delivering and collecting census forms; and
- new designs of the census form with the aim of making it easier to complete for those households most susceptible to underenumeration.

2.4 The Test provided sufficient evidence that a postback method of collecting census forms was the best strategic option. Enumerators will deliver the forms in person and they will then be used to follow-up those households who have not returned forms by post. This will provide flexibility, allowing enumerators to be redirected to follow-up in areas where response is poorest. Further, the test proved that a form designed on the basis of a page-per-person was the most efficient. (ONS, 1998)

2.5 In addition, the population definitions for the census - defined as the population which it is intended to count - have been reviewed. In particular, full information will only be collected about residents from forms completed at their usual address. In previous censuses, information has also been collected from visitors to an address on census night, to permit a count of the population 'present' on census night. Collecting information from some people twice - once where the person lives and once where they are on census night - places an unnecessary burden on the public and may confuse them. It is also possible that this could lead to an undercount of

residents if a visitor only completes a census form at their census night address, believing they have fulfilled their obligation. There was some evidence of this in the 1991 Census - the CVS estimated a figure of 100,000 people living in Britain but only enumerated as visitors (Heady *et al.*, 1994). Bearing in mind the propensity of the CVS to underestimate, the true figure may well have been higher.

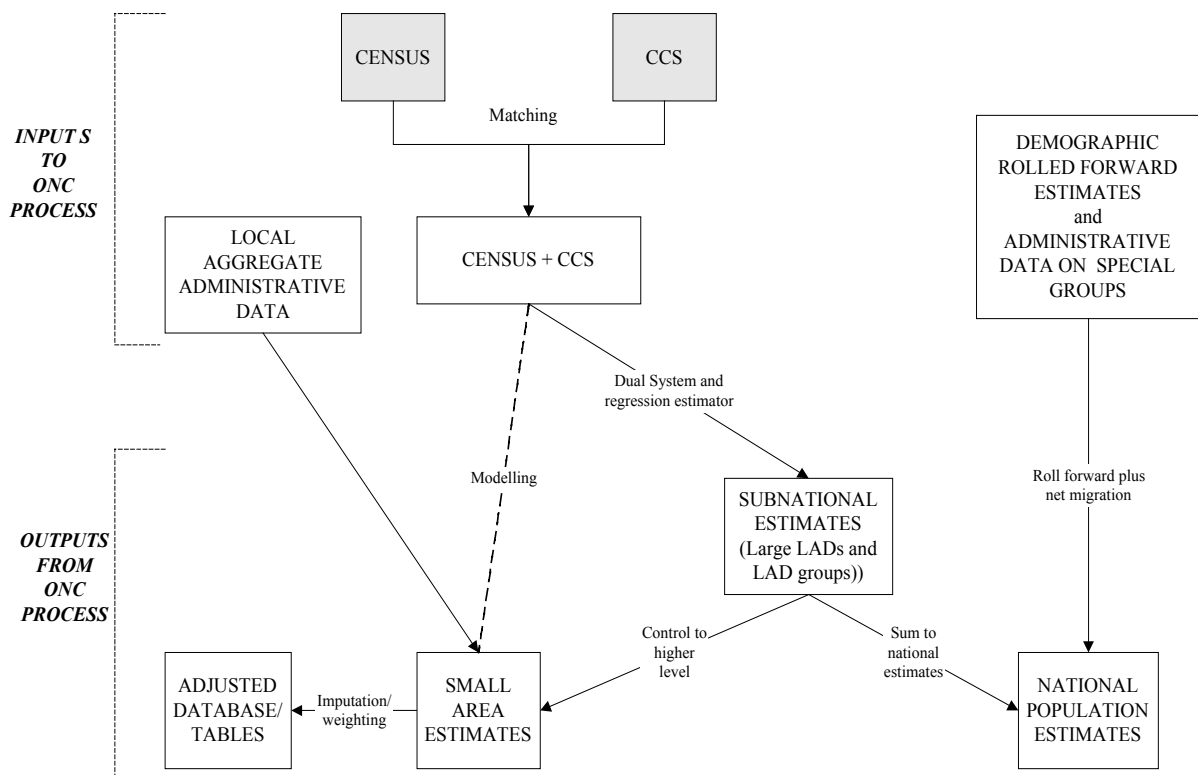
2.6 The definitions of residence for subgroups of the population with more than one address - such as students, armed forces and long-distance commuters - have been reviewed. A principle has been adopted that the residence should in general be the address at which the person spends (or intends to spend) the majority of their time.

2.7 Publicity for the Census will also be used in maximising coverage. The precise strategy has yet to be established but a key element will be the need to target those groups of the population most susceptible to underenumeration - ethnic minorities, young men and the elderly. Liaison with local community groups will play a key role in this.

### Overview of the One Number Census methodology

2.8 A schematic representation of the ONC process is presented in Figure 1.

**Figure 1. A schematic overview of the One Number Census process**



## ***Census Coverage Survey***

2.9 The key element in the ONC methodology is the Census Coverage Survey (CCS). This will be a large, highly focused survey designed solely to investigate the coverage of the census. Unlike previous censuses, this post-enumeration survey will not measure the quality of answers to census questions. This is being determined during the testing and planning stages for the Census.

2.10 The CCS will involve a complete re-count of all households and people in a sample of postcode units. Interviewers will collect information believed to be associated with underenumeration from all households and residents within households in the selected postcodes, by completing a short questionnaire for each household. The precise sample size of the CCS has yet to be decided but it is likely to be in the range of 250,000 to 600,000 households. It is unlikely that any one survey organisation will be able to provide the numbers of interviewers required (perhaps as many as 10,000). Therefore, the most likely strategy will be, in the main, to re-use the best Census enumerators, albeit in a different area to that they enumerated in the Census itself.

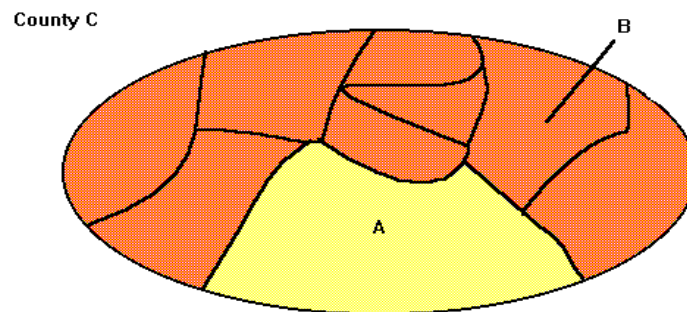
2.11 Research has been carried out into the practicalities of undertaking the CCS. Following the 1997 Census Test a pilot CCS was undertaken in the Brent area of London. Findings from this initial work suggest that in 2001 the CCS fieldwork will begin 3-4 weeks after census day and will take 3 weeks. Interviewers will work in pairs which will make conducting the interviews easier (one asking questions, one writing answers) and will also allow interviews to be carried out later in the evening, if necessary, to maximise coverage. The questionnaire will include those questions that are thought to be associated with underenumeration and those necessary for matching the data to that from the census. Further work is planned to develop the practical aspects of undertaking a postcode based CCS taking forward the lessons learnt. The CCS will be part of the Dress Rehearsal for the 2001 Census.

## ***The production of sub-national (pseudo-county) census-based population estimates***

2.12 Although it would be ideal to design the CCS to provide direct estimates of the undercount for each local authority district (LAD), this would lead to a very large sample size. Therefore, the survey will be designed for groups of LADs - so called *pseudo-counties*. Some large LADs may be separately stratified - those with high populations, such as Birmingham and Glasgow. This is illustrated in Figure 2. The total number of pseudo-counties in England and Wales is likely to be around 50.

2.14 Enumeration districts (EDs) or output areas (OAs) in Scotland will be stratified into groups by a 'Hard to Count' (HtC) index since it is expected that underenumeration will, at a local level, be higher in certain areas characterised by particular social, economic and demographic characteristics. This will be a national index from 1 (easiest to count) to 5 (hardest to count). The components of the HtC index will be 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).

**Figure 2. Illustration of counties and pseudo ‘counties’. County C consists of 9 LADs of which A is particularly large and therefore it is treated as a separate stratum in the design. The other LADs within county C are grouped into a pseudo ‘county’ B.**



2.15 The selection of the postcodes for the CCS sample will be a two stage process:

- A number of EDs will be selected from each pseudo-county, ensuring that **all** LADs are represented in the CCS.
- A fixed number of postcodes will be selected from each ED.

2.16 The data from the CCS will be matched with those obtained from the census to estimate the number of people who were missed from the census and their characteristics. This combined census and CCS information will then be used to produce census-based estimates of the population for the pseudo-counties by age and sex using an innovative combination of standard techniques for estimating population totals from surveys. This procedure will involve a method known as capture-recapture which allows an estimate to be made of the number of people missed from both the census and the CCS.

2.17 Details of the proposed methodology for designing the CCS are given in Annex B. Estimation procedures are discussed in Annexes B and C.

### ***Producing the National level population estimates***

2.17 The census-based population estimates for the pseudo-counties will be summed to produce a national census-based estimate which will be compared with the national demographic estimate of the population (the estimate of the population rolled forward from the previous census). Annex D summarises work undertaken to produce the demographic estimates which will be required for a ONC. Comparisons will also be made at national, and pseudo-county levels (where possible) between the census-based estimates and aggregate level administrative records to check that specific groups such as the elderly and young children have been accurately counted.

## ***Producing local authority and Small area population estimates***

2.18 To produce census-based population estimates for all local authorities and smaller areas adjusted for underenumeration, models will be developed using the sampled postcodes in the CCS that will link the CCS data with the observed census characteristics. These will then be used to estimate the missing people in the non-sampled postcodes. These models will estimate the number of people missed in enumerated households and in missed households for each postcode by age and sex and other household and individual characteristics. The estimates produced for the local authority districts and smaller areas will be controlled to the higher-level agreed estimates for pseudo-counties. Further details of the methodology for modelling to local authority districts and smaller areas are given in Annex E.

2.19 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 weight the database, this is discussed briefly in Annex E. This last stage would allow all statistics based on the 2001 Census to aggregate to ‘One Number’.

### **3. Management and Consultation arrangements**

3.1 It is important that users of census data have confidence in the figures produced from the ONC process. Consultation plays a key role in the development, and acceptability to users, of the methodology.

#### **Timing of the ONC process and availability of census statistics**

3.2 An important concern is whether a ONC will result in a significant delay to the availability of statistics from the census. While there is likely to be some delay, preliminary discussions with census users responsible for allocating central government funds to local and health authorities have indicated a willingness to wait for the ONC results, provided it can be demonstrated that the procedure is feasible and that any delay is limited.

3.3 It is not possible at this stage to estimate when the first Census results will be published. A number of uncertainties remain. However, current estimates are that the ONC operational process will take about a year after the basic census and CCS databases are made available. The basic databases, in this context, are taken to be computer files of data captured from Census forms and CCS interviews. These will be largely unedited files. The ONC process involves:

- ‘marking’ the Census data for inconsistent answers - such as three year old married people;
- matching the Census and CCS records to obtain estimates of those missed by the Census;
- imputing answers to certain questions on the combined Census+CCS database where inconsistent answers remain;

- producing the pseudo-county estimates;
- summing to provide national census-based population estimates;
- comparing the census-based estimates with previous census-based estimates and administrative records at the national and (where possible) pseudo-county levels;
- modelling to produce census-based population estimates for all local authorities and adjusted census counts for small areas
- imputation of missing people and households or production of weights, and matching with other census variables (editing if necessary).

3.4 It should be noted, however, that even if a full ONC process was not implemented, a number of the above processes would still be necessary - those concerned with editing, imputation of missing data and the estimation of the undercount for local authority districts.

### **Organisation and the Steering Committee**

3.5 The methodological work for the ONC is currently being undertaken by a joint ONS/Academic team, under the direction of Professor Ian Diamond at the University of Southampton. The Project Board, chaired by Tim Jones, Director Of Methods and Quality Division at ONS reports to a Steering Committee which oversees the methodological development. The Steering Committee includes representatives of the academic and local authority communities, a senior representative of the Australian Bureau of Statistics and ONS directors. The members of the **Steering Committee** are:

*Chair:*

**Dr John Fox (Chair)**, Group Director, Census, Population and Health Group, ONS

*External members:*

**Dr Jim Cuthbert**, Consultant, formerly Government Statistical Service

**Professor Denise Lievesley**, Consultant, formerly ESRC/University of Essex

**Professor Mike Murphy**, London School of Economics

**Mr Tim Skinner**, Australian Bureau of Statistics

**Professor Mike Titterington**, Glasgow University

**Mr Steve Turner**, Tees Valley Joint Strategy Unit

*Office for National Statistics members:*

**Mr Julian Calder**, Group Director, Survey and Statistical Services Group

**Mr Graham Jones**, Director of Census

**Mr Tim Jones**, Director of Methods and Quality and Chair of the ONC Project Board

**Ms Judith Walton**, Director of Population and Vital Statistics

**Dr Marie Cruddas (Secretary)**, Statistician, Census Division

4.4 This consultation paper will be distributed to all members of the Census Advisory Groups and other interested parties. Any feedback and comments on the proposed methodology should be either raised at the Census Advisory group meetings in April/May, at the ONC Workshop in Leeds (see below) or by writing directly to ONS (address at the front of this paper) by **1 June 1998**.

4.3 A One Number Census Workshop is being held from 12-13 May 1998 in Leeds as part of a Joint Workshop on Census Issues. Census users will be able to raise points on the methodology with members of the ONC Project Team.

4.4 A final agreed strategy will be tested in the Census Dress Rehearsal to be held in the Spring of 1999. The key elements will be the development of the practical aspects of the fieldwork for the CCS, data capture, matching of Census and CCS responses, and piloting of the estimation and modelling processes.

## **ANNEX A. Undercoverage in the 1991 Census**

*This annex outlines factors which may have contributed to the undercount in the 1991 Census*

### **Experience in 1991**

Although the level of coverage in the 1991 Census was high, nearly 98%, there was a problem with differential undercount whereby the underenumeration was not evenly distributed throughout the population (OPCS, 1994). The population sub-groups that suffered most from underenumeration were:

- Young adults aged 20-29 (6% nationally, 11% in cities and as high as 23% for young males in some cities);
- Those in converted or shared accommodation (11% net underenumeration of households);
- Infants under one year old (estimated at 3% nationally);
- Armed forces personnel and their dependants;
- Elderly women (around 6% for 85s and over); and
- Ethnic minorities in city areas.

The following factors may well have contributed to this undercount:

#### ***Students***

The census was conducted during holiday time for some but not all students. This, together with the fact that they were enumerated as resident at their parents address, led to uncertainty over completion of the census form and resulted in some students being missed completely.

#### ***Resident six-month rule***

The rules adopted in 1991 implied that a person could only be treated as a resident at an address if they had been living there for at least six months. It is thought that this rather strict definition led to the exclusion of some of the elderly population who had recently moved into nursing and residential homes.

#### ***Contacting households***

Buildings designed for single-occupancy but converted to contain a number of individual flats led to undercounting as enumerators did not realise that additional households occupied the dwelling. Purpose-built flats with entry phones also caused problems as contact could be difficult to establish.

Due to changes in society, it was more difficult in the 1991 Census than in the past for enumerators to make contact with someone in the household. For example, there was an increase in the number of one person households and households where both people were working.

## Annex B. Design of the Census Coverage Survey

James Brown, Ray Chambers, Ian Diamond and Lisa Buckner

*This annex describes the design for the Census Coverage Survey (CCS) to follow the 2001 Census. A model based approach is adopted for the design and direct estimation. This allows full advantage to be taken of the highly correlated auxiliary information available after the 2001 Census.*

*At the time this work was carried out the working assumption was that the CCS would be designed at the county level (or using groups of smaller counties). Since then it has been decided that ‘pseudo-counties’, as described in the overview in Chapter 2, will be used. This does not affect the principles outlined below but should be borne in mind when reading this annex.*

### B1. Introduction

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 during the testing and planning stages for the census. 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 group<sup>2</sup>. 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

---

<sup>2</sup> 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.

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)<sup>3</sup> of less than  $\alpha$  per cent where  $\alpha$  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:

$c$  = 1... $C$  design level county groups in England & Wales.

$d$  = 1...5 HtC categories of postcodes.

$e$  = 1... $M_{dc}$  enumeration districts in HtC category  $d$  of group  $c$ .

$i$  = 1... $N_{dc}$  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...24 age-sex groups (0-4, 5-9, 10-14, ..., 40-44, 45-79, 80-84, 85+).

---

<sup>3</sup> RSE (also called coefficient of variation) =  $\frac{\sqrt{\text{var}(T)}}{T} \times 100$

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.

## B2. 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:

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

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

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 B5.

## B3. 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.

#### B4. 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  $\alpha_d$  and  $\beta_d$  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.

#### B5. 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. 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

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+. These are age groups that might be expected to be least well counted (based on experience in 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  $j^{\text{th}}$  component score for the  $e^{\text{th}}$  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 cent<sup>4</sup>. 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. Sample allocation for the first stage sample in Hampshire**

Index Group	Population Size	Number of Size Strata	Sample Size	Outlying <sup>b</sup>
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 <sup>a</sup>	28	-	28	-
<b>TOTAL</b>	<b>3229</b>	<b>150</b>	<b>336</b>	<b>11</b>

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.

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.

<sup>4</sup> An RSE of one per cent for total T translates into an approximate 95 per cent confidence interval on T of  $\pm 2$  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.

### **B6. 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. 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

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.

## **B7. 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)<sup>5</sup>. The distribution of enumeration districts by HtC index is given in Table 4.

**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

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

<sup>5</sup> The numbers given are approximate for confidentiality reasons.

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 B5) 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. Sample allocation for the first stage sample**

Index Group	Population Size	Number of Size Strata	Sample Size	Outliers <sup>b</sup>
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 <sup>a</sup>	2	-	2	-
<b>TOTAL</b>	<b>930</b>	<b>70</b>	<b>79</b>	<b>6</b>

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 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.

**Table 6. Mean Relative Standard Errors for 1000 simulated CCSs**

Males				Females			
Age Group	Number of CCSs	Design RSE	Average <sup>b</sup> Estimated RSE	Age Group	Number of CCSs	Design RSE	Average <sup>b</sup> 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 <sup>a</sup>	7.67	2.18 (0.978)	80-84	997 <sup>a</sup>	6.24	1.67 (0.588)
85+	721 <sup>a</sup>	10.43	3.71 (1.661)	85+	999 <sup>a</sup>	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.

The simulation shows that for a perfect CCS the proposed design in conjunction with the regression estimator performs well. In Annex C the more realistic situation of dependence between the census and CCS with CCS non-response is examined in detail for the regression estimator model.

## **Annex C. Dependence, non-response and underenumeration estimation methods**

**James Brown, Ray Chambers, Ian Diamond and Lisa Buckner**

*In this annex the estimation method described in Annex B is extended to allow for non-response in the CCS by incorporating Dual System Estimation techniques (alternatively known as capture-recapture) into the regression estimator. The performance of the new estimator is examined using simulations with varying degrees of dependency between the census and CCS counts.*

### **C1. Introduction**

The theory underlying use of the regression estimator in the CCS design described in Annex B 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 US<sup>6</sup>. 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 underestimate 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

---

<sup>6</sup> 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.

(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 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. **The triple system estimation approach has been examined as part of the ONC research but because no administrative list of sufficient quality at the individual level was available, and potential concerns over confidentiality, this approach will not be used in the ONC process.**

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.

The following section describes simulations which investigate the effects of varying degrees of dependence between the census and the CCS for different ‘Y’ counts in the regression estimator.

- $Y_{MAX}$  = 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 for a postcode.

Section C2 describes how dependence is simulated and examines the performance of the regression estimator based on  $Y_{UNION}$ , for various levels of CCS response and dependency. In C3 the three ‘Y’ counts are compared when there are various levels of dependency and a CCS response rate of 80%.

## C2. Simulation study of the impact of correlated non-response on the CCS design

The simulation described in Section B7 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 B7 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 this annex 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:

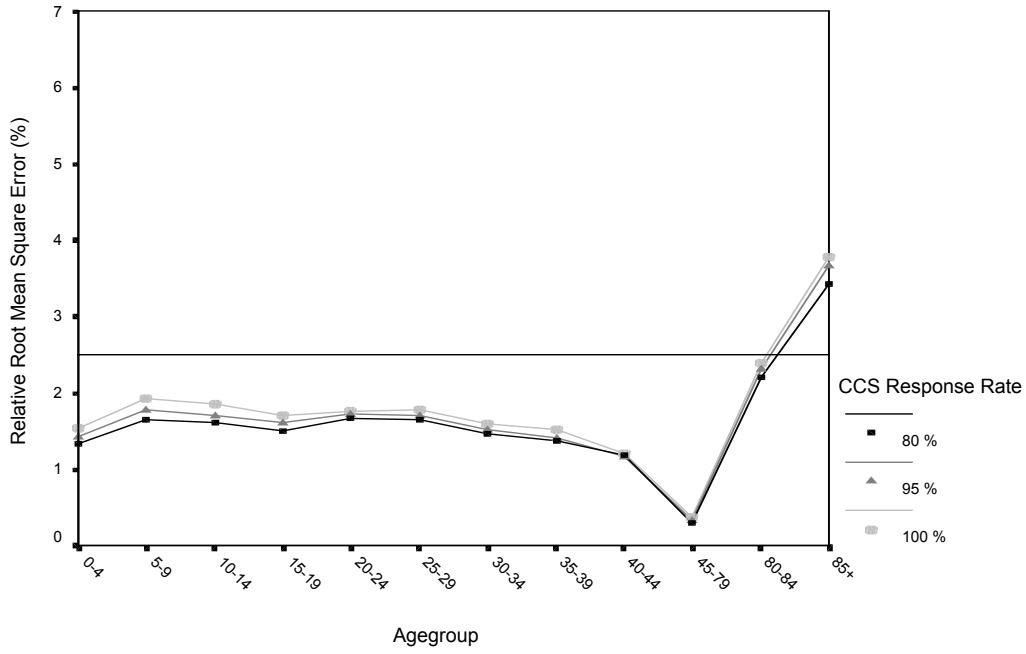
$$\text{RRMSE} = \frac{1}{\text{truth}} \times \sqrt{\frac{1}{1000} \times \sum_{j=1}^{1000} (\text{observed}_j - \text{truth})^2} \quad (5)$$

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

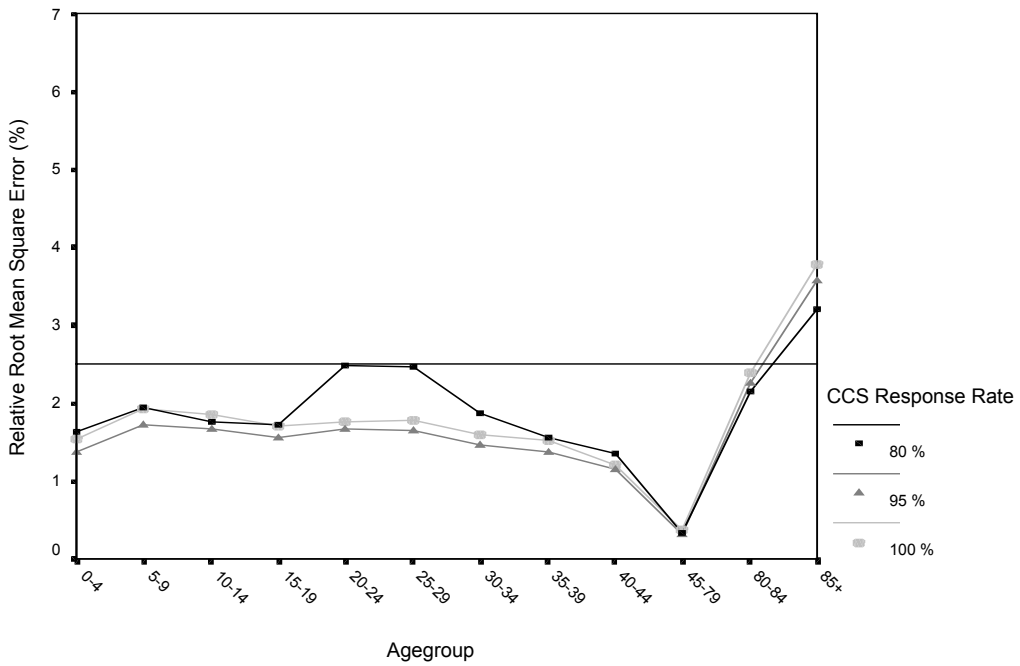
$$\text{Relative Bias} = \frac{1}{\text{truth}} \times \frac{1}{1000} \times \sum_{j=1}^{1000} (\text{observed}_j - \text{truth}) \quad (6)$$

for the same groups. The results are presented in a series of graphs for varying odds ratios by sex. Figures 3 to 5 are for males and show the RRMSE.

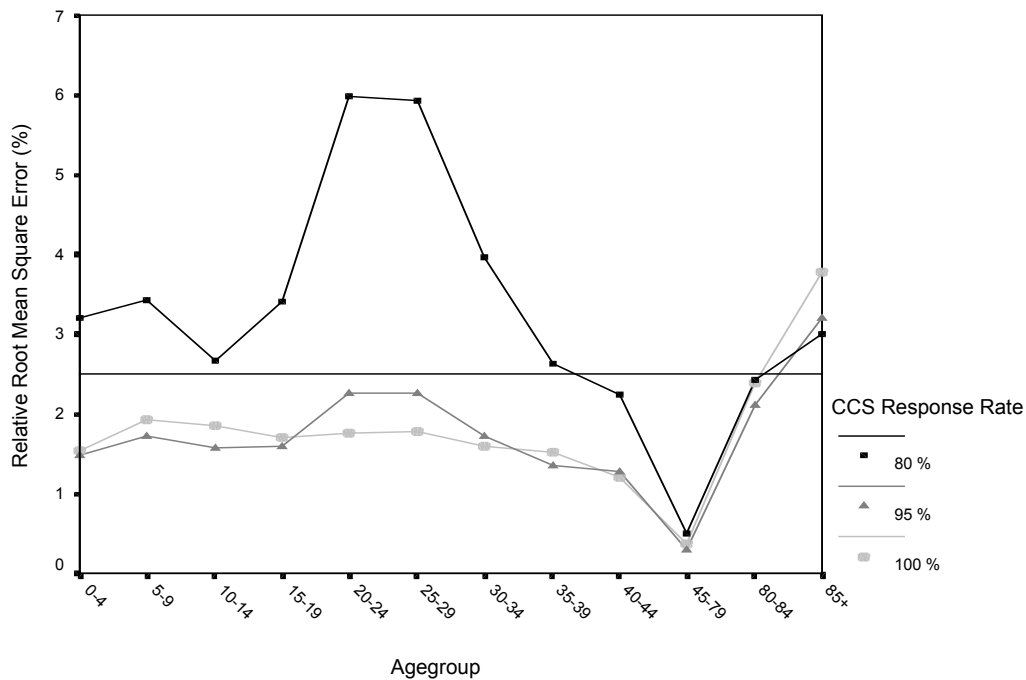
Figures 3 and 4 show that for odds ratios of 0.1 and one the RRMSE remains below 2.5 per cent even when the CCS response rate is 80 per cent. However, as the odds ratio increases above one the same people tend to be missed by the census and the CCS. In this case, as the CCS response rate falls the RRMSE 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 co-operate 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.



**Figure 3. Performance of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 0.1**

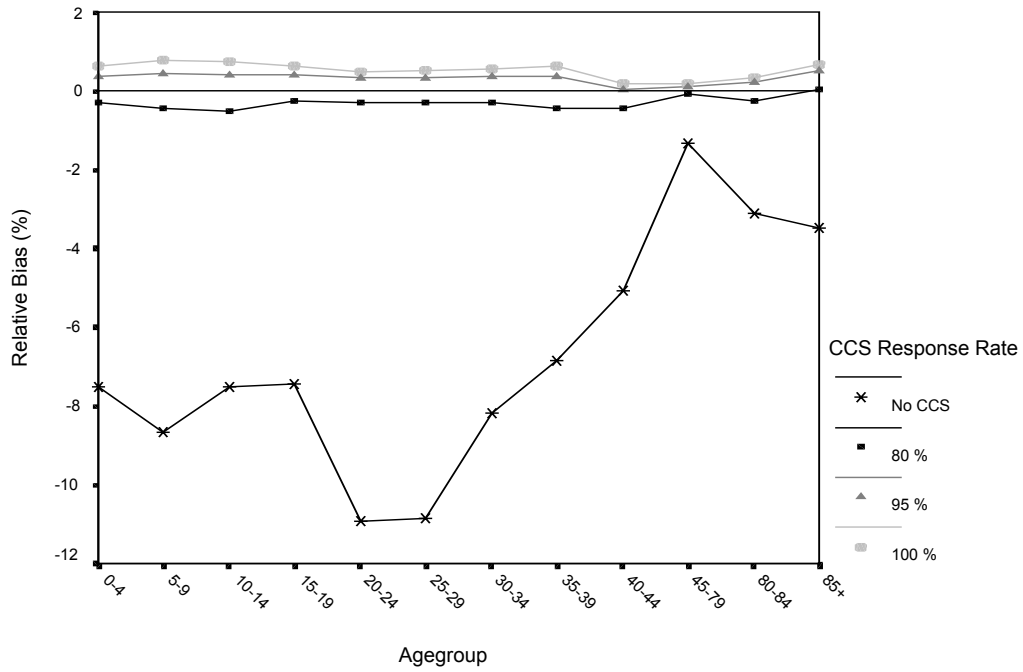


**Figure 4. Performance of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 1.0**

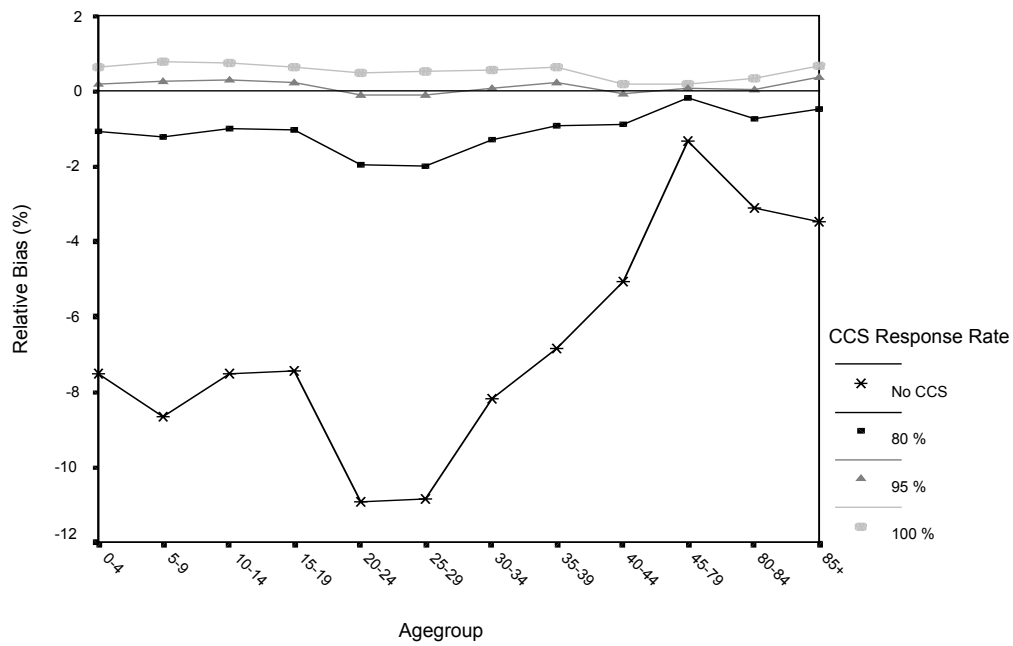


**Figure 5. Performance of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 10**

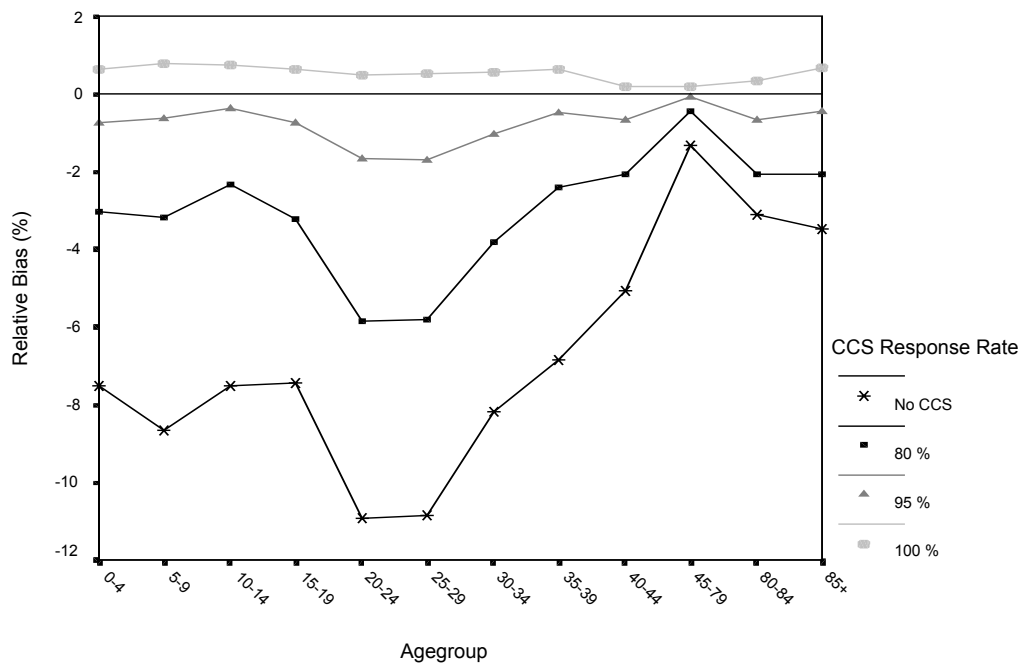
Figures 6 to 8 present the relative biases for males for the changing odds ratios. For reference the relative bias for the unadjusted census counts is also presented (termed ‘No CCS’ in Figures 5 to 8). Figures 6 to 8 show that as the RRMSE increases with the odds ratio the negative bias of the regression 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 5 and 8 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.



**Figure 6. Relative bias of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 0.1**



**Figure 7. Relative bias of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 1.0**



**Figure 8. Relative bias of adjusted county totals for males by Census Coverage Survey (CCS) response rate: odds ratio = 10**

**C3. The impact of correlated non-response in the CCS for different 'Y' counts.**

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 for 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 9 shows the full set of graphs for the RRMSE and relative bias for males. Figure 9 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.

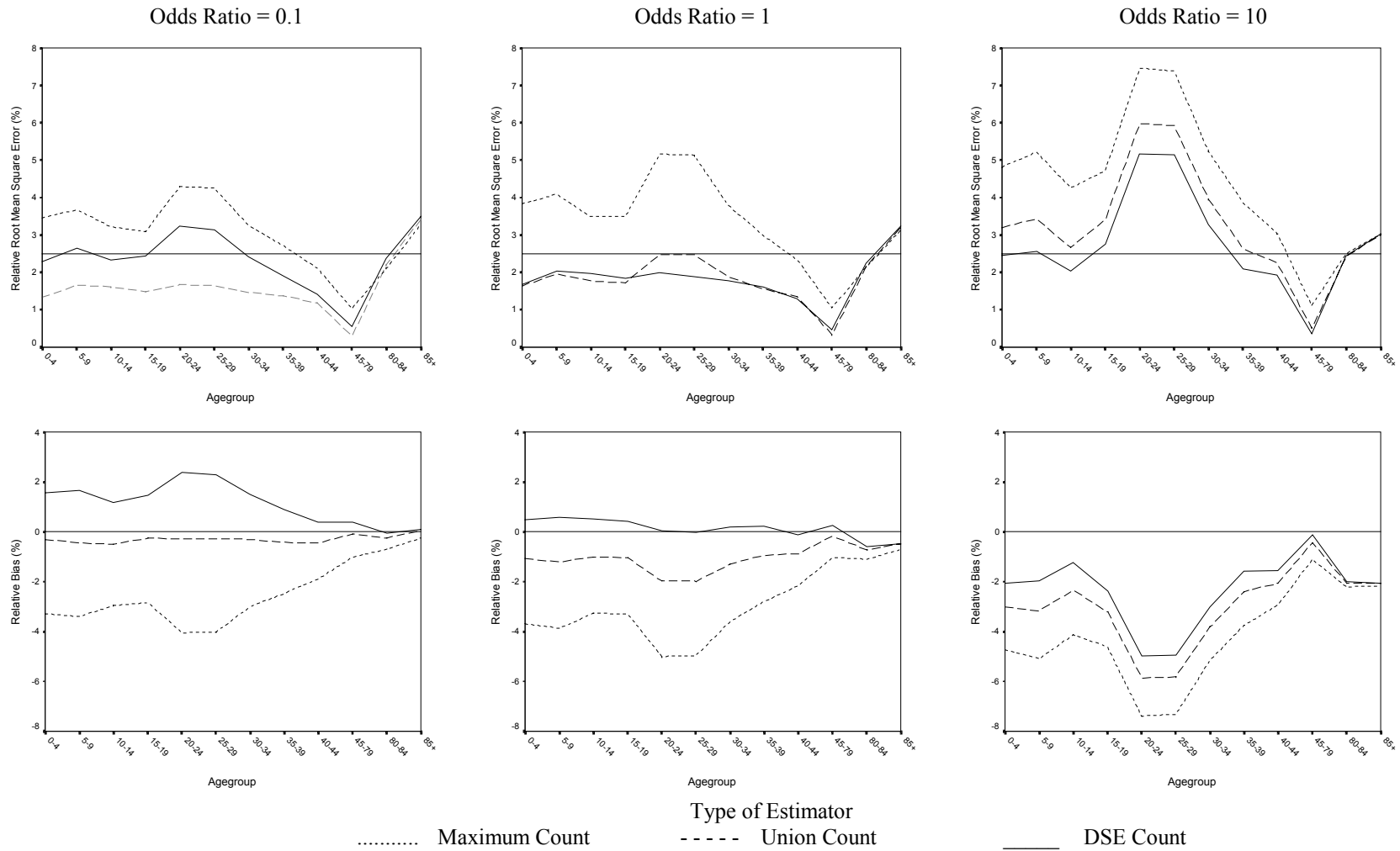


Figure 9. Performance of adjusted county totals for males by type of count used in the regression estimator for varying odds ratios: CCS response rate = 80 %

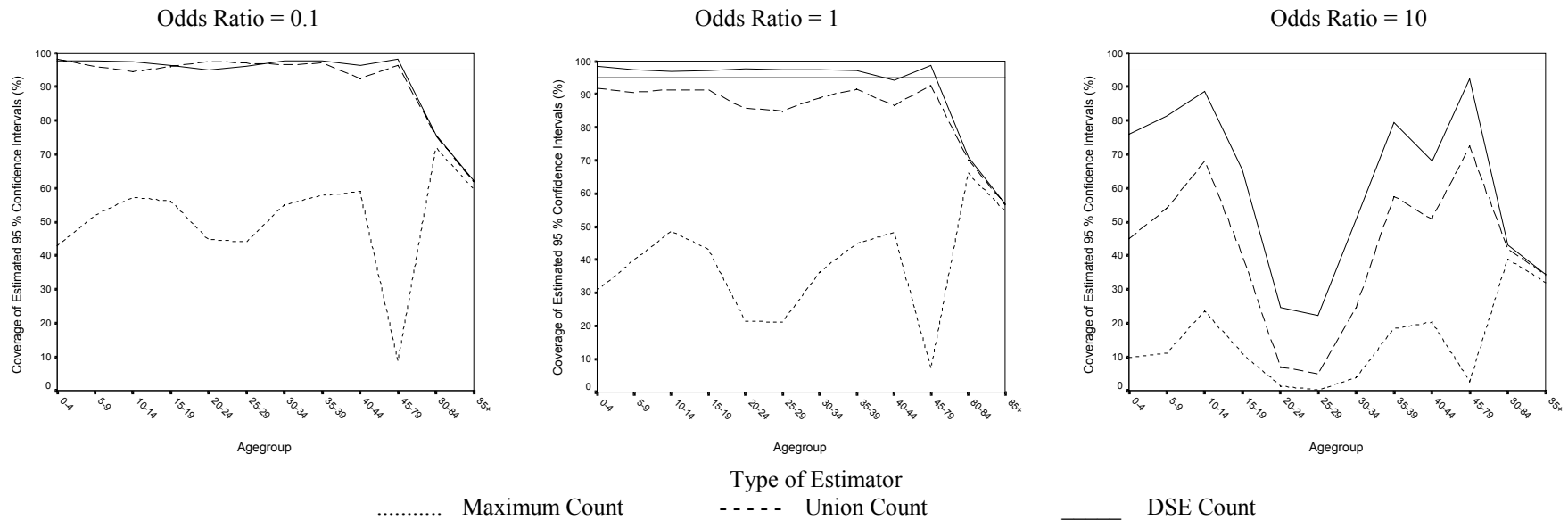


Figure 10. Performance of the adjusted county totals variance estimator for males by type of count used in the regression estimator for varying odds ratios: CCS response rate = 80 %

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 10 presents graphs to show coverage of the ultimate cluster variance estimator for each of the estimators for the range of odds ratios.

Figure 10 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.

## **Annex D Demographic analyses in support of a One Number Census**

**Roma Chappell, John Charlton and Ian Diamond**

*Annex D summarises work undertaken to produce the demographic estimates which will be required for a One Number Census.*

### **D1. Introduction**

Annual mid-year population estimates are produced for England and Wales by the Office for National Statistics, and for Scotland and Northern Ireland by the General Register Office (Scotland) and the Northern Ireland Statistics Research Agency (NISRA) respectively. These demographic based estimates are produced by rolling forward the most recent census counts, allowing for births, deaths and net migration since census day. An allowance is made for underenumeration in the census base. The estimates are available at the national level for each of England, Wales, Scotland and Northern Ireland, and at sub-national levels which, in England and Wales, are the local authority districts, Unitary Authorities and London boroughs.

The demographic analyses being carried out in support of a One Number Census (ONC) has three high level aims:

- (a) To provide a plausible range for the national population numbers in 2001, by age and sex, as benchmark for the 2001 Census data.
- (b) To establish whether the 1981 Census was the most appropriate point from which to produce rolled-forward population estimates.
- (c) To examine the extent to which sub-national demographic estimates might be used as part of the ONC strategy.

This Annex reports progress for England and Wales on the work that has been done so far towards meeting these aims and describes work that is planned for future.

### **D2. A plausible range for national population estimates in 2001**

The first stage of the ONC strategy compares census results to the demographic based estimates at the national level. If the census results are within a plausible range of the demographic estimates then the census will be confirmed as the “Gold Standard”. The plausible range for the demographic estimates is an error band taking account of sources of error in the rolled forward estimates. Some errors can be quantified e.g. sampling errors on the International Passenger Survey flows, but quantifying other sources of error will involve a degree of judgement.

Sources of errors in the mid-year estimates are under constant review by Population and Vital Statistics Division in ONS. The annual mid-year population estimates are produced using the best data and methodology available at the time.

The sources of error in the annual mid-year national population estimates arise from the census base or from the components of the estimates which are the birth and death data (natural change), migration data and data for the armed forces resident in the UK.

### ***The census base***

Each decennial census is liable to general under-enumeration and miscounting of particular population sub-groups such as armed forces personnel, elderly people and babies. As outlined in this paragraph, allowances for this underenumeration have been made in the past and the ONC strategy is building that experience in to the planning for 2001. The issues for each of the groups listed above are discussed in a little more detail:

- The very young - Birth registration data have traditionally been used to produce population estimates of the under ones in census year rather than relying on the census. The potential use of child benefit statistics as a check on counts of children in 2001 will be researched.
- The elderly - After the 1991 Census it was found, by comparing census counts to DSS retirement pension data, that the very elderly were undercounted in the census. The comparison showed that DSS and census counts were in agreement up to the age of 79 but an enhancement of 63 thousand was needed to the resident count for persons aged 80 and over.
- The armed forces - For the compilation of the demographic estimates armed forces data are collected from the Defence Analytical Services Agency annually. These data are in respect of where HM armed forces personnel are stationed rather than where they reside. The Population and Vital Statistics Division of ONS are researching the feasibility of improving the quality of the estimates of HM armed forces on a residence basis as part of the ongoing work towards improving the quality of the annual mid-year population estimates. Data in respect of foreign armed forces personnel and their dependents resident in England and Wales are collected on a residence basis annually.

More detail on the accuracy of the census base is given in the next section of this report which specifically addresses the issue of which census should form the base for the rolled forward estimates.

### ***Natural Change***

The natural change component of the estimates (births minus deaths) is widely accepted to be reliable - compulsory registration systems have been in force in the country since 1839 and improvements have been introduced over time.

### ***Migration***

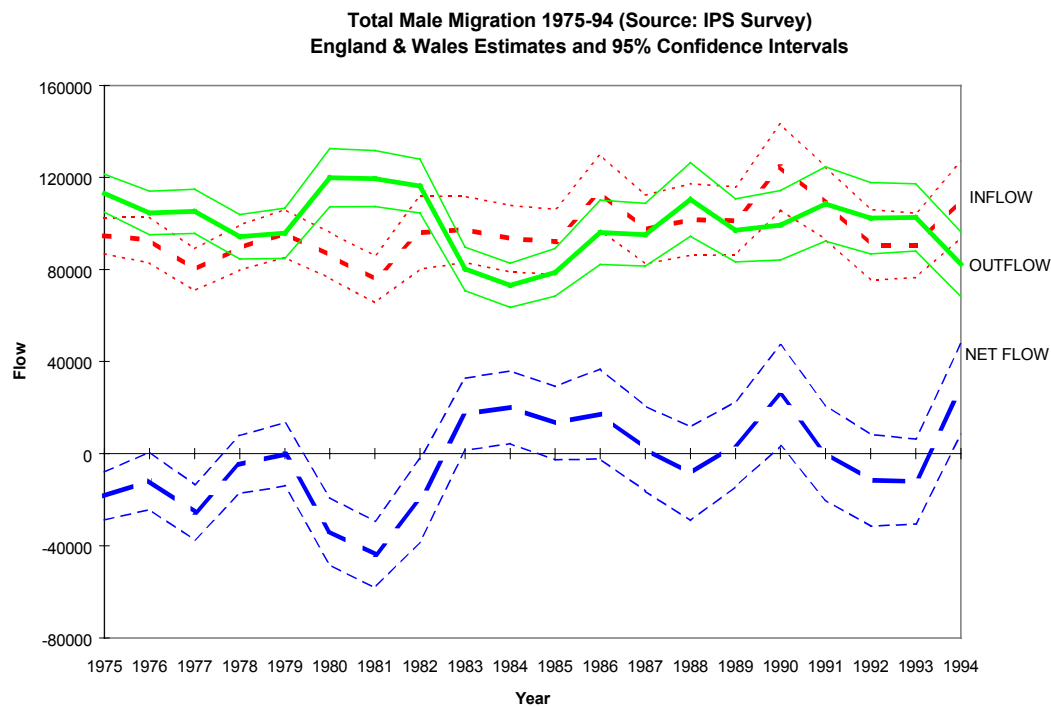
There are two types of migration, international migration which is to and from the UK and internal migration which is movements within the UK. Internal migration does not affect the accuracy of population estimates at the national level.

In the compilation of population estimates the estimated migration refers to the civilian population only, as armed forces personnel are identified separately rather than being included as part of the rolling forward process.

The main source of the international migration data is the International Passenger Survey (IPS) which is a voluntary survey which aims to collect data on movements to and from the UK by all travellers. Since international migrants are a relatively small proportion of all travellers the migrant sample size is small, about 1,400 immigrants and 750 emigrants in 1995. Thus the sampling error is high with confidence intervals of around  $\pm 10\%$  for each of the in- and out- migration flows for a single year.

The IPS produces estimates of the number of international migrants to and from England and Wales each year, based on stated intentions at the time of interview. Figure 11 shows the estimates with 95 per cent confidence intervals for males, for all ages combined (the graph for females is similar). The standard errors have been calculated based on the sampling fractions, and allowing for a survey design effect of 1.2. The sampling error increased from the early 1980s when the sampling fraction was reduced.

**Figure 11.**



If the benchmark population estimates for the 2001 Census are rolled forward from 1981, then 20 years worth of IPS data will be used in the estimates for 2001. So the confidence intervals for migration over this period will be smaller in relative terms than in each year. The pooled data have been used to produce Figures 12 and 13 shown later in this annex in section D3.

### ***Estimation of non-sampling errors for the IPS***

As in any survey in addition to sampling errors, there will also be non-sampling errors in the IPS. The main sources of bias to the estimates of migrant flows are:

- for a number of reasons, people interviewed on arrival may not give an accurate response to the question on how long they intend to stay.
- it is felt that emigrants from England and Wales are more likely to be able to conduct the interview without any language problems than immigrants to England and Wales
- the IPS only monitors arrivals and departures during the daytime and the assumption is made that those arriving on night flights are similar.

The ONS Social Survey Division carry out the IPS interviews and they are being consulted about the possibility of making an allowance for these sources of error.

### ***Other sources of error in international migration***

The estimates of migration between the Irish Republic and the UK are derived currently from the Irish Labour Force Survey (LFS) and the National Health Service Central Register (NHSCR). Ways to estimate the migration flows between the UK and the Irish Republic were investigated by the Population Statistics Division of ONS in conjunction with the Central Statistics Office in Dublin. An improved method will be used from mid-1997 onwards. The flow from the Irish Republic to GB, will be calculated from the Irish LFS, the NHSCR for England and Wales and the Irish Country of Residence Survey. This will be calculated by statisticians in the Irish CSO and agreed with the ONS. For 2001, estimates of error based on sampling errors of the Irish LFS and including an allowance for non-sampling errors will be researched in co-operation with the migration statistics unit at ONS.

The IPS can not cover all migration, e.g. only a few asylum seekers could be picked up by the IPS and people who originally enter the country as visitors would not be defined as migrants by the IPS. The Home Office supplies data to ONS annually, in respect of asylum seekers and these “visitor switchers”. A visitor switcher is a person entering the UK who is admitted as a short-term visitor and then stays for a year or longer.

### **D3. Establishing the census base for demographic based estimates**

The census base used for rolling forward to produce intercensal mid-year population estimates has to be adjusted for under-enumeration in the census. In the past this has been achieved mainly through the census post enumeration survey, which in 1991 was the Census Validation Survey (CVS). However the 1991 CVS failed to detect the full extent of the under-enumeration. Comparisons of 1991 Census results with the population estimates rolled forward from 1981 and other demographic analyses suggested a net undercount of about 1.2 million residents. In the final analysis the population estimates rolled forward from 1981 were deemed to be more reliable at the national level than the census and are thus also the basis of estimates from 1991 onwards.

The 1981 post enumeration survey was the most thorough that had been carried out up until that time. As well as the survey itself some use was made of administrative records in arriving at the final estimate of the undercount as 240,000 persons (about 0.4%). Since, potentially, the 1981 post enumeration survey could have suffered from a similar problem

to the 1991 CVS, there is a need to confirm that the 1981 Census provides the best base for rolled-forward national population estimates. This has been done by producing demographic based population estimates, for comparison with the census counts, using the cohort analyses methodology.

### ***Cohort Analyses***

Cohort Analyses are a way of producing population estimates which are independent of censuses. Starting with births in each year from an initial year, the birth cohorts are aged on, subtracting deaths and allowing for net migration, to give an estimate of the numbers who remain. This approach is similar to that used to produce a rolled forward estimate except that it is not linked to any census. Instead a much longer series of births, deaths and, crucially, migration is needed. The cohort analyses undertaken for the One Number Census are based on all births from 1911 onwards.

### ***Comparing cohort estimates with 1971, 1981 and 1991 Censuses and national population estimates***

Comparisons have been made between:

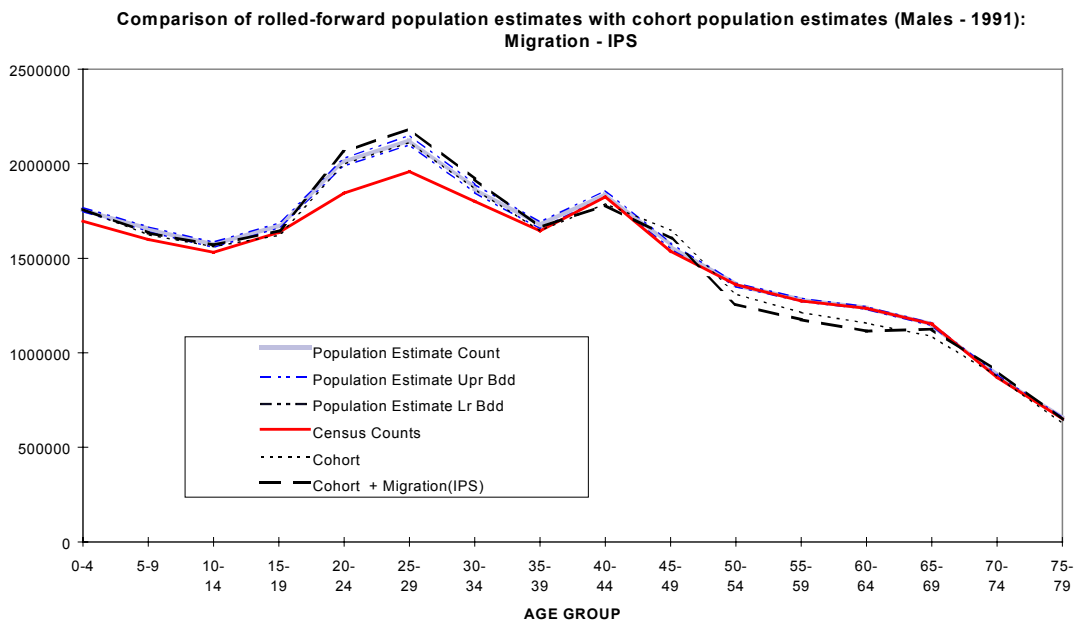
- cohort analysis results, without taking into account the effect of migration;
- cohort analysis results adjusted for migration effects;
- unadjusted census data for 1971, 1981, and 1991;
- official population estimates.

Figures 12 to 15 show, for 1991, and 1981 the comparison of the cohort approach with census data and population estimates, with and without allowance for the effects of migration. Figures 12 and 13 compare the estimates for 1991, and also show 95% confidence intervals based on IPS sampling errors. Points to note are the difference between the census counts (data not adjusted for under-enumeration) and the Population estimates (the *revised final* estimates for 1991). This equals (by definition) the under-enumeration that was assessed for 1991 when the population estimates were made. Also, in Figure 12 a census undercount of males in their early twenties is clearly visible when compared with the cohort estimates. The cohort estimates are higher for this age group than the official estimates, but this was not the case in 1971 or 1981. Figures 14 and 15 compare the estimates for 1981. Once again it is clear that population estimates produced by both methods are similar. Similar results were obtained for 1971. The difference between census results and both types of population estimates is much smaller than in 1991. This gives an indication of the likely greater accuracy of the 1981 Census. There were more people found in the census than the cohort analyses (adjusted for migration) predicted for ages 30-49 in 1971, 40-59 in 1981, and 50-69 in 1991. This could be because the migration figures failed to include a large number of immigrants in the 1960s, when there was a large net inflow (Coleman and Salt, 1992). This will be investigated further using the Labour Force Survey.

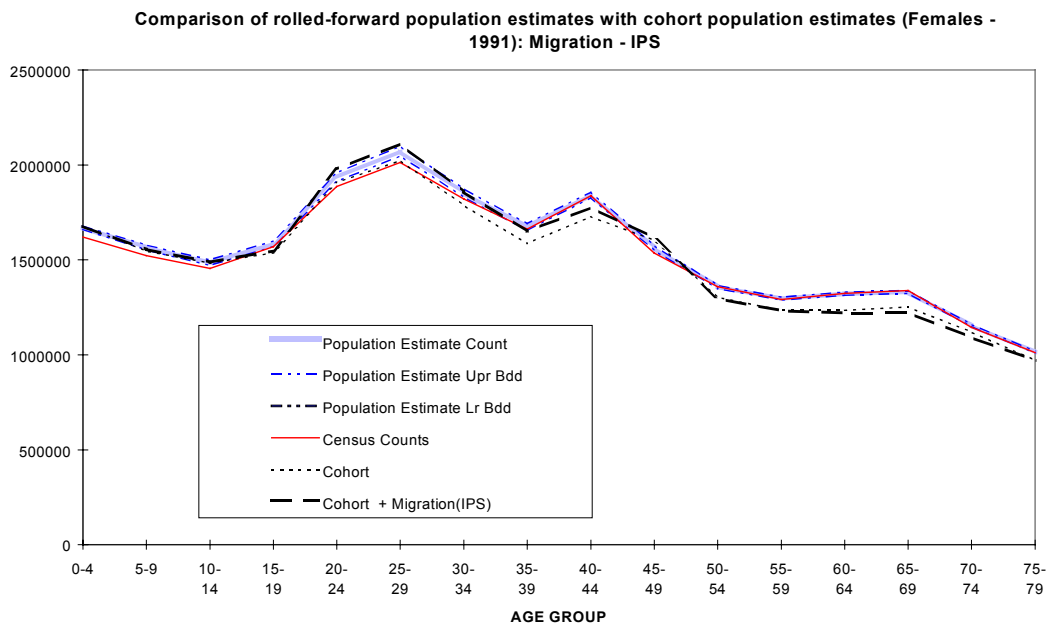
A cohort approach is independent of censuses whereas the official population estimates are essentially based on a census. However the cohort approach requires migration estimates spanning a greater number of years. Results from the cohort analyses provide population estimates closely in line with 1971 and 1981 Census data, and most similar to the official population estimates for 1981. This corroborates the official estimates at the national level,

and suggests that 1981 official population estimates provide the best base from which to roll forward population estimates for the 2001 Census benchmark.

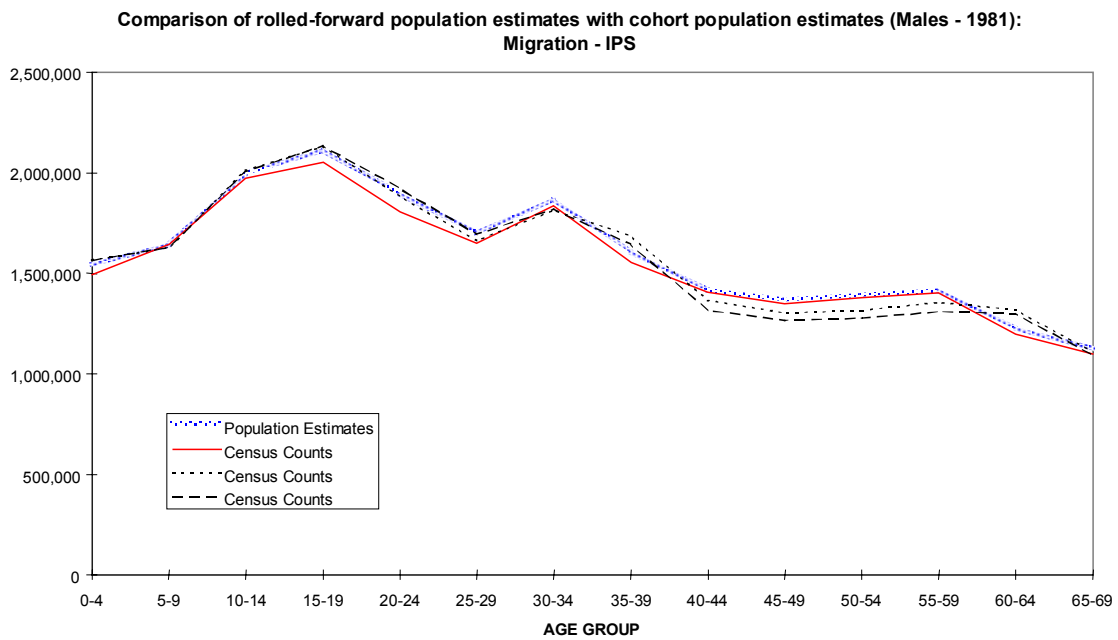
**Figure 12.**



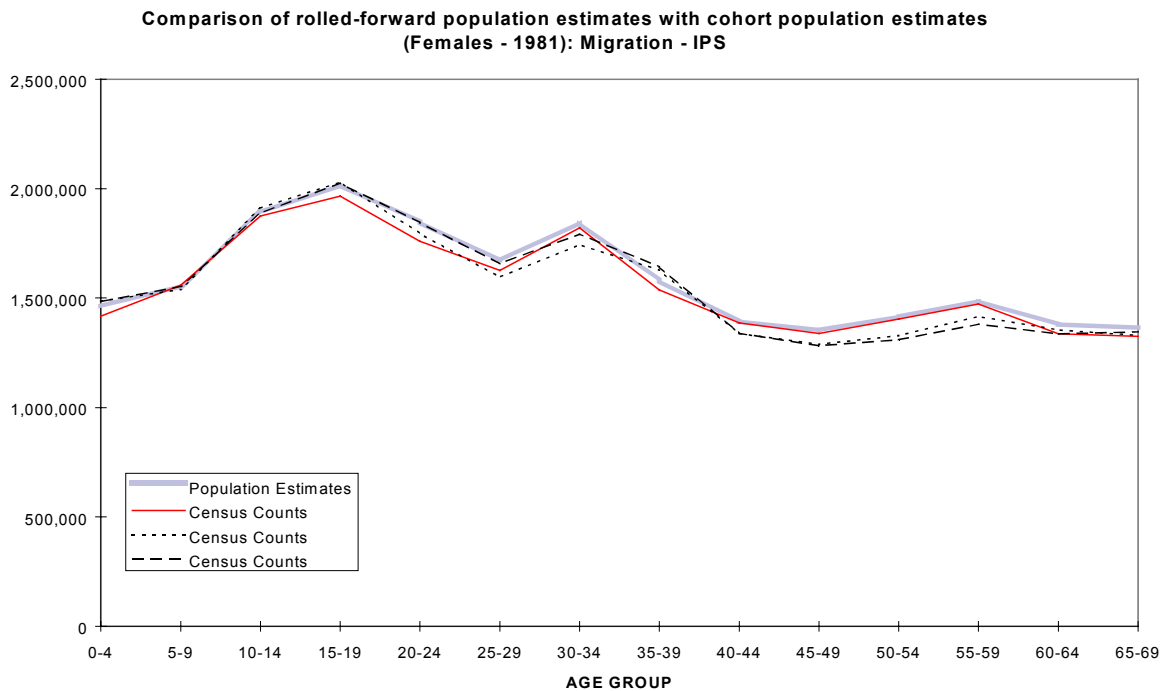
**Figure 13**



**Figure 14.**



**Figure 15**



***An alternative check on which census provides the best basis for rolled-forward population estimates - Analysis of census non-respondents in the Longitudinal Study***

The ONS Longitudinal Study (LS) consists of linked census and vital registration data on a one per cent sample of the resident population of England and Wales. Selection into the LS is by birth date, and the study was designed as a continuous, multi-cohort study, with

subsequent samples being drawn at each census, using the same selection criteria, and linked into the dataset (Hattersley and Creeser 1995). It includes the 1971, 1981 and 1991 Censuses, and there are no adjustments for under-enumeration. Each year, new members are entered by virtue of being born on LS dates or by immigration (if born on LS dates) and exited by death or emigration. LS members are also traced to the National Health Service Central Register (NHSCR). The data on an LS member include everything collected in the censuses.

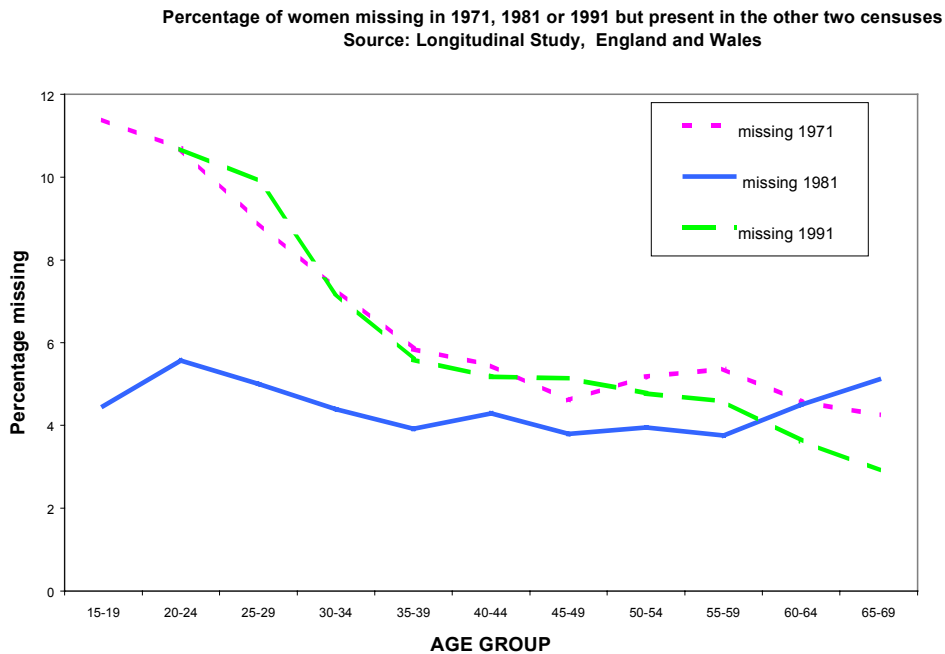
The LS can be used to identify the types of people who are apparently missing from censuses, analogous to the Reverse Record Check used by Statistics Canada (Burgess, 1988). There are some caveats, however:

- those missing may merely be not linked, e.g. because they have the wrong date of birth on a census form. Birth dates are more likely to be correct on the NHSCR than on the census. The census form is usually filled in for the household by one individual, who may record the date incorrectly;
- those missing may have been temporarily out of the country as part of an absent household on census night;
- the LS does not cover Scotland, so those who move there will not be included in the linkage;
- the NHSCR records are only amended when people re-register with a different GP or de-register, for example when leaving England and Wales. Young healthy men may not always re-register promptly with their GP after moving, and those leaving the country will probably not de-register with the NHS.
- When an individual is absent from a census no data are available at that point in time. Information on age, sex, and country of birth will be available from other censuses, however.

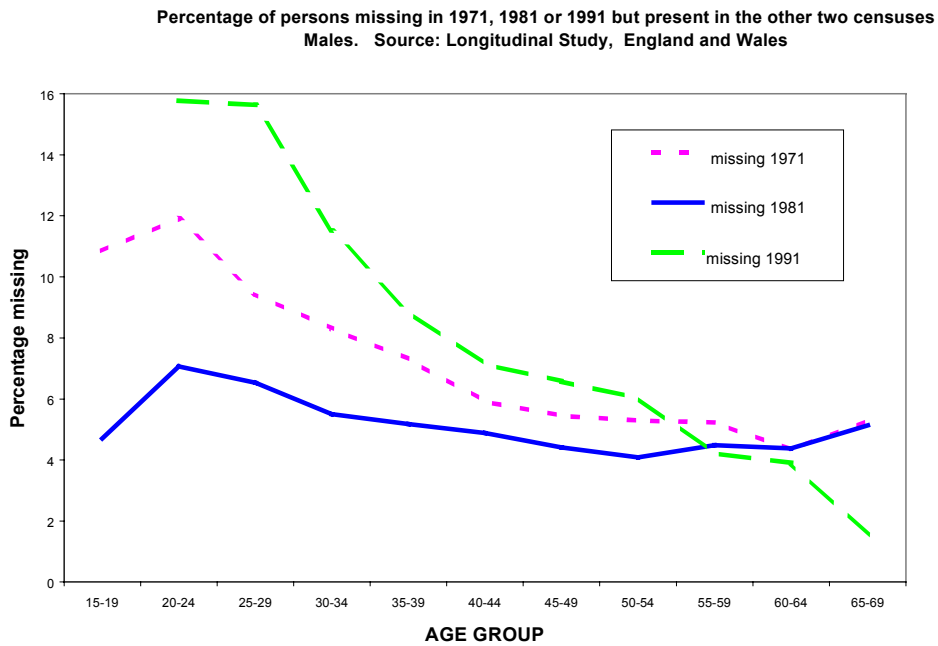
Whilst bearing these caveats in mind, Figures 16 and 17 show estimates of those missing in 1971, 1981, or 1991 but present in the other censuses. These data exclude those who died or were known to have immigrated. Only individuals traced at NHSCR in 1981 and 1991 were included. LS “undercounts” are higher than the undercounts achieved in the censuses, partly because “missing” could also be due to non-linkage of records. Of the three censuses, 1981 had the smallest proportions of missing individuals. Also, young men (and to a lesser extent young women) aged 20-30 were most prone to be missing in the LS. This was the group suspected of being most prone to under-enumeration in the 1991 Census. The results also give a feel for the relative magnitude of the under-enumeration problem in 1991.

We have considered whether it would be possible to use the LS to help to estimate the types of people missed by the census in 2001. However a verified LS database it is only likely to be available one and a half years after 2001. Our provisional view is that any potential benefits of using the LS would be unlikely to warrant making a case for a speedier link.

**Figure 16.**



**Figure 17.**



#### **D4. Sub-national analyses**

The question that needs to be addressed is what work needs to be extended to sub-national levels. What geography is feasible - regions, counties, local authority districts or groupings of local authority districts? Would it be possible to do cohort analyses at this lower geographical level or produce error bands to this level? If possible would the results be of sufficient quality to be used?

The potential usefulness of a cohort analysis at sub-national level is limited by the quality of the migration data. In making a population estimate migration is the most difficult component to get right. At present therefore it is recommended that cohort analyses are not pursued at the sub-national level. There is an ONS project within the Population and Vital Statistics Division that is examining the potential use of Family Health Service Registers to make estimates of migration at the county and sub-county level. This project will report at the end of June 1998 on the feasibility of using FHSA registers. However, even if this were thought to be feasible there would be difficulty in obtaining these data for all the past years that would be required for producing a cohort analysis.

The first test of the validity of the census and Coverage Survey will be based on comparing national data against demographic estimates. Even if the figures agree there is still the need to validate the census results for smaller geographical areas. Rolled-forward population estimates and administrative sources may be used in combination as part of this procedure. However the error bands on the non-census figures at local levels will be much wider than those at the national level, because there is more migration between smaller areas than large ones which is difficult to measure. An assessment will be made of the size of these errors, but there will be no attempt to assess the errors each local authority individually. Rather a model will be developed which takes into account the size and type of local authorities and past problems with enumeration, and assessments will be made for broad age/sex bands.

## **D5. Conclusions**

The cohort analyses confirm that the adjusted 1981 Census forms an appropriate base population for estimates up until the 2001 Census. For the national level, confidence intervals for the 2001 population estimates can be calculated based on the sampling error inherent in the International Passenger Survey. For the total population these amount to about  $\pm 50,000$ . Further work is required to quantify the potential extent of non-sampling error. Quantified sampling and non-sampling error will enable plausible national population ranges to be constructed, against which the 2001 adjusted Census results can be validated. The extent to which further demographic analyses are required at a sub-national level is currently under consideration.

## Annex E. Modelling down to small areas

James Brown, Ray Chambers, Ian Diamond and Lisa Buckner

*The ultimate aim of the ONC project is a census database fully adjusted for underenumeration. This requires a procedure that estimates weights in order to adjust for underenumeration. These weights can then either be applied to the records for enumerated people and households to produce a weighted database or used in the imputation of missing people 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.*

### E1. 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 (that is pseudo-county). 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 Annex C 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<br>(age, sex, marital status, ethnicity, economic status)     | $\underline{X}$ |
| ii)  | a vector of their household characteristics<br>(tenure, building type, multiple-occupied, number of residents) | $\underline{Z}$ |
| 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 Annexes B and C, using the CCS

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.

## E2. 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 problems<sup>7</sup> the following multilevel multinomial model can then be fitted for the CCS sample postcodes:

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

where  $\alpha_r$ ,  $\underline{\beta}_r$ ,  $\underline{\gamma}_r$  and  $\eta_r$  ( $r = 1, 2$ ) are the unknown fixed parameters associated with the variables  $\underline{X}$ ,  $\underline{Z}$  and  $S$  and  $\lambda_{rlm}$ ,  $v_{rklm}$  and  $\varepsilon_{ijklm}$  are independent random error terms with  $\lambda_{rlm} \sim N(0, \sigma_{rlm}^2)$ ,  $v_{rklm} \sim N(0, \sigma_{rklm}^2)$  and  $\varepsilon_{ijklm}$  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

<sup>7</sup> When there is a choice between a census and CCS measure the CCS measure will be used in the modelling.

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.

### E3. 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  $\hat{\nu}_{rklm}$  to get predicted probabilities for each of the different types of individuals and households in all sampled areas given by:

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

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 spatial<sup>8</sup> 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  $\hat{\pi}_{0ijklm}, \hat{\pi}_{1ijklm}, \hat{\pi}_{2ijklm}$ .

### E4. 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  $klm$ .)

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

<sup>8</sup> 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.

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

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

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

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

## E5. 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 E4. 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.

## E6. Simulation Study Methodology

The same underlying method used for the CCS design simulation in Section B7 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 Annex B.

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:

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

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.

## E7. Discussion of Results

The following discussion is preliminary and only present results for the overall adjustment described in Section E6, 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.

**Table 7. Root Mean Square Error across all simulations for Enumeration District totals**

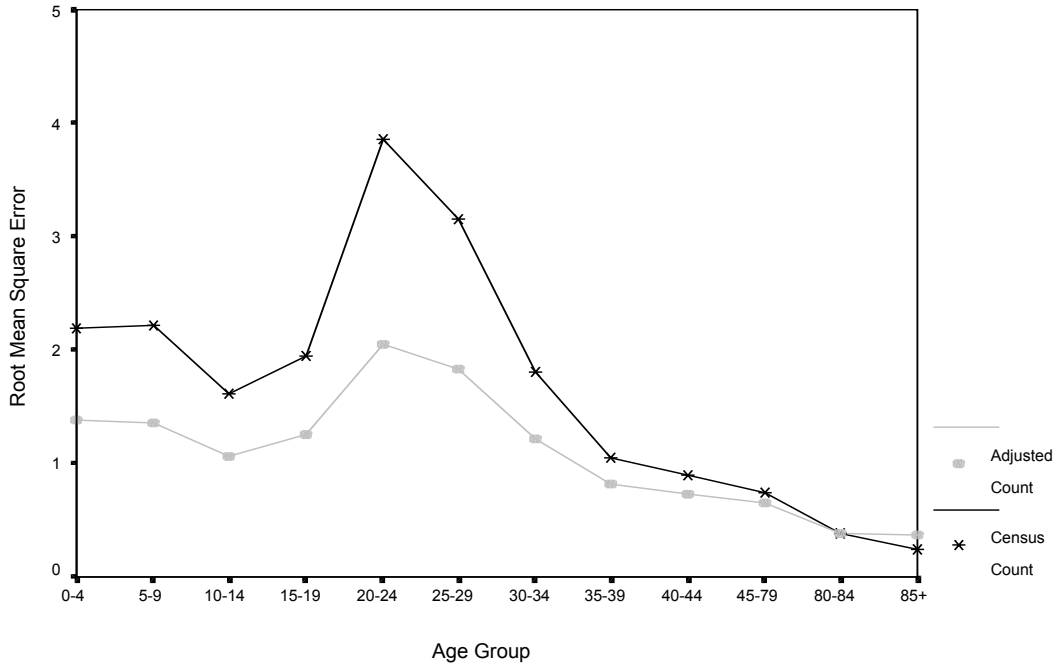
	HtC Index				
	Very easy	Easy	Medium	Hard	Very hard
Census count	9.94	12.09	13.23	15.92	23.89
Adjusted count	3.64	3.90	4.25	4.42	5.84

**Table 8. Bias across all simulations for enumeration district totals**

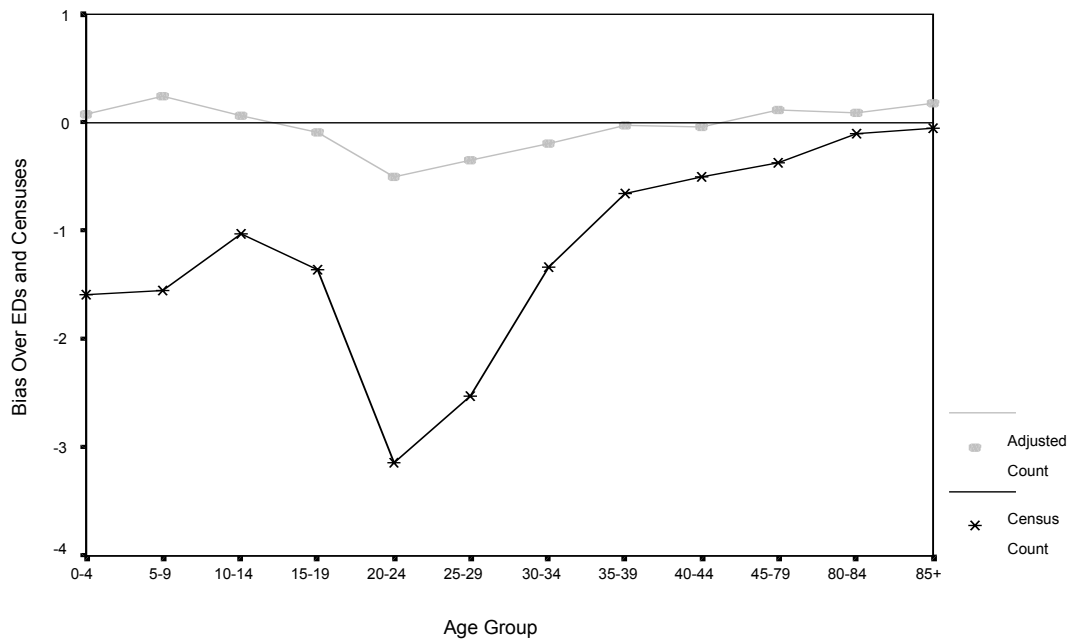
	HtC index				
	Very easy	Easy	Medium	Hard	Very hard
Census count	-9.25	-11.22	-12.29	-14.86	-21.84
Adjusted count	-0.08	-0.66	-0.65	-0.01	-0.90

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 18 and 19 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.



**Figure 18. Performance of adjusted enumeration district totals for males relative to unadjusted census totals: HtC = 5 ('very hard')**



**Figure 19. Bias of adjusted enumeration district totals for males relative to unadjusted census totals: HtC = 5 ('very hard')**

Figure 18 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 19 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.

## 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.
- Burgess, R.** (1988). Evaluation of reverse record check estimates of undercoverage in the Canadian census of population. *Survey Methodology*, **14**: 137-156.
- Coleman, D. and Salt, J.** (1992) *The British Population*. Oxford University Press, Oxford.
- 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.
- Hattersley, L. and Creeser R.** (1995). Longitudinal Study 1971-1991. History, organisation and quality of data. LS no 7. OPCS/ HMSO, London
- 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** (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** (1998). Evaluation of the main objectives of the 1997 GB Census Test. *Census Advisory Group Paper* (98)01.
- 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.