

ONE NUMBER CENSUS STEERING COMMITTEE**Statistical models to estimate underenumeration in the 2001 Census of England and Wales**

1. This paper was presented by James Brown at the Population Association of American Conference in Chicago, 2-4 April 1998.
2. Work has taken place to extend the simulations on the small area modelling to include another explanatory variable (Primary Activity Last Week) in addition to age-sex and hard to count index. An additional random effect at the enumeration district level has also been added. The inclusion of these are detailed at the start of Section 5.
3. **The Steering Committee are asked to:**
 - a) **note the paper and the progress made;**
 - b) **provide any comments at the meeting on the 27 April 1998, or in writing by 10 May 1998.**

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

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

March 1998

**STATISTICAL MODELS TO ESTIMATE UNDERENUMERATION
IN THE 2001 CENSUS OF ENGLAND AND WALES**

BY
J. J. Brown⁽¹⁾
I. D. Diamond⁽¹⁾
R. L. Chambers⁽¹⁾
L. J. Buckner⁽²⁾

**University of Southampton
Department of Social Statistics
Highfield
Southampton
Hampshire
SO17 1BJ
UK**

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

Tel: +44 (0)1703 592527

+44 (0)1329 813507

Fax: +44(0)1703 593846

+44 (0)1329 813532

Email: j.j.brown@soton.ac.uk

lisa.buckner@ons.gov.uk

ABSTRACT

As a result of lessons learnt from the 1991 Census of England and Wales, minimisation of underenumeration has been given high priority in the research agenda for the 2001 Census. In recognition, however, that 100 per cent coverage will never be achieved, the One Number Census (ONC) project was established to measure underenumeration in the 2001 Census and, if possible, fully adjust the statistics from the census for that undercount. This paper describes the methodology that is being put forward for this purpose. It is based on estimation of total underenumeration using a Census Coverage Survey (CCS), together with a statistical modeling approach to estimate differential underenumeration. In particular, spatially smoothed random effects models are suggested for micro level adjustments to the census database. Results from a simulation study show that the proposed methodology works in a simplified case. Extensions to more complex situations are discussed.

Keywords: CENSUS UNDERENUMERATION; CENSUS COVERAGE SURVEY; SAMPLING; REGRESSION ESTIMATION; MULTINOMIAL MODELS; RANDOM EFFECTS

1. INTRODUCTION

Every ten years there is a census in the UK, the last one was in 1991. This was in fact three national censuses carried out on the same day; one for England and Wales, one for Scotland, and one for Northern Ireland. Within England and Wales the census provides national figures on the population. The census also reports at the county level, (55 counties in 1991), the local authority district level (403 in 1991), ward and enumeration district level (about 110,000 in 1991). Of these the local authority counts are used by parliament to distribute local authority funding. In addition there is a lower level, the postcode. A postcode is a small group of address points (mean around 15) set up by the post office for the delivery of mail. The census does not report counts for postcodes although their location within enumeration districts is reported. Some postcodes cross enumeration district boundaries. There are 1.6 million postcodes in the UK; the majority of these are in England and Wales

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

England and Wales, Scotland and Northern Ireland in rebasing the population estimates as well as for the census users in interpreting census counts.

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

The priority for the development of the 2001 Census is to ensure that the maximum possible coverage is achieved, and in particular that the differential nature of any underenumeration is minimised. However, despite efforts to maximise census coverage, it is only realistic to expect there to be some degree of underenumeration. The One Number Census (ONC) project aims to measure this level of underenumeration in the most acceptable way, to provide a much clearer link between the census counts and the population estimates, and if possible to adjust all the census counts (which means the individual level database itself) for underenumeration. All counts will then add to 'One Number'. This has entailed a re-think of the design of the post enumeration survey and how this can be used to measure underenumeration at the different levels of aggregation reported by the census. Work is also being done on its integration with other measures of underenumeration provided by administrative records and demographic analysis which is not presented in this paper.

In Section 2 of this paper we give an overview of the whole ONC process. In Section 3 we present the proposed design for the CCS to measure underenumeration at a sub-national level. Section 4 of the paper introduces a model for estimating underenumeration at the smallest census reporting level and suggests how this could be applied to fully adjust the census database. In Section 5 a simulation study of the CCS design is carried out firstly at the aggregate level and then for the small area model. In Section 6 we draw some conclusions and discuss further work that is needed.

2. OVERVIEW OF THE ONC METHODOLOGY

The process outlined below represents the proposed methodology for a ONC as part of the 2001 Census in England and Wales. However, this is subject to change and refinement as the main consultation with users is about to take place. Extensions of the work to include Scotland and Northern Ireland are also being considered. The ONC process can best be considered as consisting of the four main stages summarised below and illustrated in Figure 1 at the end of this paper. The CCS inputs to the process are discussed in more detail in Sections 3 and 4.

The first two stages will be to produce the best estimate of the population by age and sex at national and local authority district level. (For the smaller local authority districts it will be

necessary to group contiguous districts.) The sub-national level estimates will then be aggregated to produce a national census-based estimate and compared with a demographic estimate of the national population rolled forward from the previous census. Charlton *et al.* (1997) summarise work underway to optimise the methodology used to produce the demographic estimates. These stages are essential for rebasing the population estimates.

The third stage will produce estimates for lower levels of geography and for other characteristics of people and households. The final stage is either to impute records for households that are estimated to have been missed and people estimated to have been missed from counted households or to produce a database weighted for underenumeration. This last stage would allow all statistics based on the 2001 Census to aggregate to 'One Number'. These last two stages have less relevance for central government distribution of resources but are relevant for local authorities that need to allocate resources at a local level. In addition, it is of course important that users know the implications of any undercount on figures for small population groups and geographical areas.

3. CENSUS COVERAGE SURVEY

The CVS, which followed the 1991 UK Census, 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 CCS in 2001 should address coverage exclusively with the aim of facilitating the estimation of underenumeration at a sub-national level (by age and sex); and to allocate this underenumeration down to small areas. Information on the quality of census data will be obtained from the question testing programme, the 1997 Census Test and possibly through a survey carried out after the Census Dress Rehearsal in 1999. This allows for a much shorter doorstep questionnaire. Savings in time can be translated into a larger sample size.

The proposal is for a postcode-unit-based survey. This requires the re-enumeration of a sample of postcode units rather than households. This clustering also helps to enable a larger sample size. While that does not necessarily improve the direct estimation of underenumeration due to the increase in variance as a result of correlation between households within postcodes, it is important for estimating adjustments at lower levels.

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

¹ Each design level group is either a single local authority district or group of smaller contiguous districts. 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.

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

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

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

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

X_{aiedc} = 2001 unadjusted census count.

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

where:

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

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

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

$i = 1 \dots N_{dc}$ postcodes in HtC category d of group c of which n_{dc} are in the sample S_{dc} , the rest are in the non-sample R_{dc} .

$a = 1 \dots 24$ age-sex groups (0-4, 5-9, 10-14, ..., 40-44, 45-79, 80-84, 85+).

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

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

3.1 Stage one of the CCS design

A robust approach to design for stage one of the CCS assumes a stratified homogeneous super-population model for the distribution of true 2001 population 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 3.4.

3.2 Stage two of the CCS design

The second stage of the CCS design consists of a random selection of postcodes within selected enumeration districts. The design is based on a selection of the same number of postcodes within each sampled enumeration district using simple random sampling without replacement. Given that size stratification and optimal allocation was used at stage one of the design, so that probability of selection of an enumeration district is approximately proportional to its size, this means that within a HtC category each postcode has approximately the same probability of inclusion in the sample.

3.3 The CCS model for estimation

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

sex groups for each HtC category. The simple regression model stratified by HtC index for an age-sex group is:

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

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

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

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

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

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

3.4 Case Study: A Prototype Stage One CCS Design for a Local Authority District in England

An anonymous local authority district in England was chosen by ONS. 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 the local authority district there are 930 enumeration districts which had a non zero population in the 1991 Census and therefore a HtC index value. (The matching to the national HtC index was done by ONS to preserve the anonymity.) The distribution of the districts by the index is given in Table 1. This shows a fairly even distribution across the categories of the HtC index.

TABLE 1
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

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 females 0-4, males 20-24, males 25-29, males 30-34, and females 85+. The choice of these key variables is based on a coverage analysis of the 1991 Census. In addition, two 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 928 districts principal components analysis was used to reduce the number of size variables. The first three component scores defined by these key size variables, which accounted for over 96 per cent of their original variability, were then used within each HtC index category to form strata by applying Ward's linkage (SAS Institute Inc., 1990) in a cluster analysis. A minimum cluster size of at least two enumeration districts was imposed. Clusters based on single enumeration districts were highlighted as outlying and included in the sample with probability one.

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

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

where P_{je} is the j^{th} component score for the e^{th} enumeration district, and V is the variance-covariance matrix of the six original size variables calculated from the 928 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 2.5 per cent³. Neyman

³ An RSE of one per cent for total T translates into an approximate 95 per cent confidence interval on T of ± 5 per cent.

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

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

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

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

From Table 2 it can be seen that in most cases the sample size required from a cluster is one and there would be little or no gain from increasing the number of clusters. For two categories the clustering has allocated enumeration districts to single district clusters. Applying other clustering algorithms in the final design may reduce this. 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 85 enumeration districts, slightly less than 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 922 enumeration districts not classified as outlying and taking a sample of 77. These ranged from 2.66 per cent for those females aged 0-4 to 10.43 per cent for those males aged 85+. The six age groups in the design variable all had expected RSEs of less than 3.3 per cent.

3.5 Conclusions on the CCS Design

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 (2). 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 the local authority deals with the practical application of the design. It shows that the theoretical framework proposed can be applied to an actual local authority district with feasible results.

4. ADJUSTING FOR UNDERENUMERATION WITHIN SMALL CENSUS AREAS

The ultimate aim of the ONC project is a single individual level census database fully adjusted for underenumeration. This requires a procedure that estimates weights at an individual level to adjust for underenumeration. These weights can then either be applied directly to produce a weighted database or used in the imputation of missing people at a very small area for both counted households and missed households. This section proposes a modelling approach to obtain these weights and briefly discusses the options regarding weighting and imputation.

4.1 Status of Individuals after the Census and CCS

Let us assume that the CCS has taken place in a sample of postcodes within each design level group. Without loss of generality only one design group is considered. For those postcodes in the sample there are two lists of individuals, one from the census and one from the CCS. These lists can, in principle, be matched to produce a single list of individuals containing all those individuals found in the census with any extras from the CCS. This is a slightly different assumption to the one in Section 3 and recognises that the CCS will not find all the people that the census does. The assumption is that no one is missed by both. This is a particularly strong assumption for some areas and work will be needed to assess the robustness of the approach with respect to this assumption.

At the individual level each person has:

- | | | | |
|------|--|-----------------|-----|
| i) | a vector of their socio-economic characteristics
(age, sex, marital status, ethnicity, economic status) | \underline{X} | |
| ii) | a vector of their household characteristics
(tenure, building type, multiple-occupied, number of residents) | \underline{Z} | (5) |
| 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;
- (6)

- 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 Section 3 with additional consideration for CCS non-response possibility through the use of administrative lists, there are gold standard age sex totals at the design group level. By gold standard it is meant that these are counts which have been adjusted for underenumeration using an 'optimal strategy'. The aim is to share the 'extra' people amongst the enumeration districts.

4.2 Multinomial model for small area adjustments

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

$$\begin{aligned} Y_{ijklm} &= 0 \text{ when individual } i \text{ of household } j \text{ is counted in the census} \\ Y_{ijklm} &= 1 \text{ when individual } i \text{ is missed in the census but his/her household } j \text{ is counted by} \\ &\quad \text{the census and the individual is counted in the CCS} \\ Y_{ijklm} &= 2 \text{ when individual } i \text{ and his/her household } j \text{ are both missed in the census and the} \\ &\quad \text{individual is counted in the CCS} \end{aligned} \tag{7}$$

Outcomes one and two are only observable in the CCS areas by matching with the CCS. Let

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

In general these probabilities will depend on the characteristics of the person, household, postcode, etc. Putting aside measurement error problems⁴ the following multilevel multinomial model can then be fitted for the CCS sample postcodes:

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

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

This is a similar approach to that taken by the US Census Bureau in 1990 to solve a slightly different problem. In their case the problem was to impute whether an individual found in their Post Enumeration Survey (PES) was counted in the census or not where matching was not conclusive. They used the predicted probability from a hierarchical

⁴ When there is a choice between a census and CCS measure the CCS measure will be used in the modeling.

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. This is detailed in Belin *et al.* (1993) along with the estimation procedures used to fit the regression model.

4.3 Prediction for non-sampled postcodes

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

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

where $r = 1, 2$ 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 random 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 spatial⁵ 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 that 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 $\hat{\pi}_{0ijklm}$, $\hat{\pi}_{1ijklm}$, $\hat{\pi}_{2ijklm}$.

4.4 Adjusting the Census

The next stage is to adjust the census counts. Let N_{ijklm} be the census count of individuals with the set of characteristics given by i from households with characteristics given by j in location klm . (E.g. employed married males aged 20-24 living in rented accommodation within location klm .)

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

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

⁵ 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.

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

4.5 Imputing extra people verses weighting

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

For the people from the missed households, there will be a set of groups of people given by the different \hat{m}_{2ijklm} . The task is then to fit the individuals back together as households. It is likely that this will require modelling to estimate the number of completely missed

households 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 4.4. The possibility of using weighted tables for census output is also being investigated as an alternative to imputation. This will also require the production of household weights for the census tables at a household rather than individual level.

5 SIMULATION STUDY OF THE CCS DESIGN AND ESTIMATION MODELS

The simulation study involves two separate simulations. The aim of the first simulation is to examine the performance of the CCS design, and particularly the gain from regression estimation, when the second stage sample is taken. The aim of the second simulation is to demonstrate that the small area model proposed in Section 4 is feasible. The anonymised individual records from the 1991 Census, augmented by the HtC index, for the local authority district used in Section 3.4 were used in both simulations. The district 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 930 enumeration districts (five have only one postcode, one has 40 postcodes, and the median is 14 postcodes)⁶. As already stated the distribution of enumeration districts by HtC index presented in Table 1 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 study was to create individual probabilities of being counted in a census. Each individual was given a fixed probability of being counted in a census based on their age, sex, and the category of their enumeration district on the HtC index. This was done by simple random sampling with replacement from the population of Estimating With Confidence enumeration district adjustment factors. These are acknowledged to give the best estimates at a small area level (Simpson *et al*, 1997) for the 1991 Census. In addition it is expected that underenumeration will vary by other characteristics other than those in the relatively simple Estimating With Confidence adjustments. In particular it was thought that employment status would be important and so an additional variable, ‘Primary Activity Last Week’, was selected. This was introduced as independent of age, sex, and HtC effects. Table 3 gives the categories for the variable as defined in the 1991 Census. The differential underenumeration factor is multiplied by an individual’s fixed probability based on their age, sex, and HtC index to calculate their fixed probability of being counted in a census based on all four variables. The values chosen reduce the probability of economically inactive individuals being counted relative to economically active individuals.

The final stage was to add a clustering effect at the enumeration district level. With the help of ONS, to preserve anonymity, each postcode in the local authority was given its Ordnance Survey location on a 100-metre grid. The grid references are given in ten metre units. This was used to locate each enumeration district approximately by taking the

⁶ The numbers given are approximate for confidentiality reasons.

average location of the postcodes within each enumeration district. Using this a distance matrix based on geographic location was calculated for the enumeration districts. Starting in the centre of the local authority district grid the enumeration districts were ordered according to the nearest enumeration district that had not yet been selected.

TABLE 3
Differential Undercount by Categories of Primary Activity Last Week Variable

Primary Activity Last Week Category	Differential Underenumeration Factor
Employee working full-time	1.0185
Employee working part-time	1.0185
Self-employed, employing others	1.0185
Self-employed, not employing others	1.0185
Government employment / training scheme	0.9500
Waiting to take / start a job	0.9500
Unemployed / looking for work	0.9000
School / full-time education	0.9500
Unable to work (illness or disability)	1.0000
Retired from paid work	1.0000
Looking after home / family	1.0000
Other economically inactive	1.0000
Aged under 16 years	1.0000

A cluster effect for each enumeration district was then created from a standard normal distribution. This was correlated with the effect for the previous enumeration district on the list such that $\rho = e^{-\ln(2) \times \sqrt{\frac{\text{distance}}{10}}}$. This meant the correlation was always positive with a maximum value of one and decreased as the distance between the enumeration districts increased. The values two and ten were chosen such that two enumeration districts at the same grid reference would have a correlation of one and two enumeration districts at adjacent grid references would have a correlation of 0.5. Further work will certainly need to involve varying this correlation structure. Figure 2 shows this decreasing correlation as distance increases.

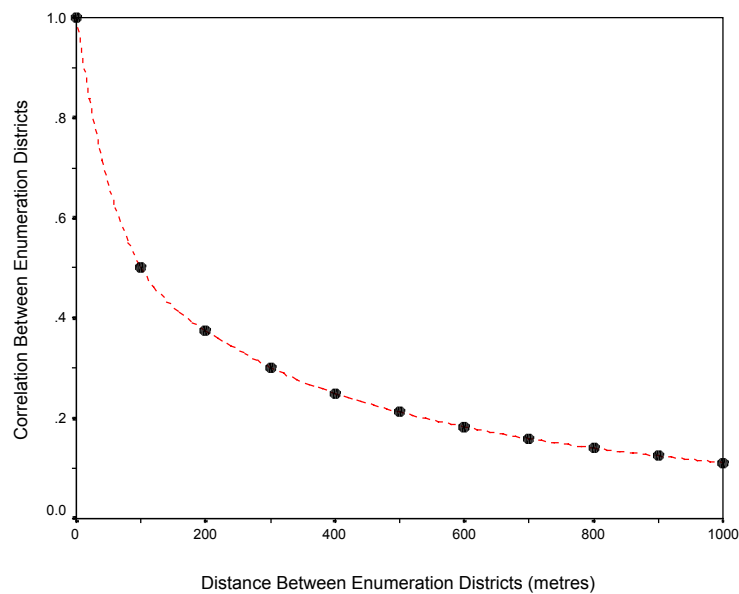


Figure 2. The relationship between correlation of the cluster effect with the neighbouring enumeration district and the distance from the neighbouring enumeration district.

These enumeration district effects were then multiplied by 0.1 and added to each individual’s fixed probability. This was done so that values from the standard normal distribution would not swamp the fixed probability. Any probabilities less than zero were set to zero and any probabilities greater than one were set to one. The distribution of these individual probabilities is shown in Figure 3 for males from HtC category five.

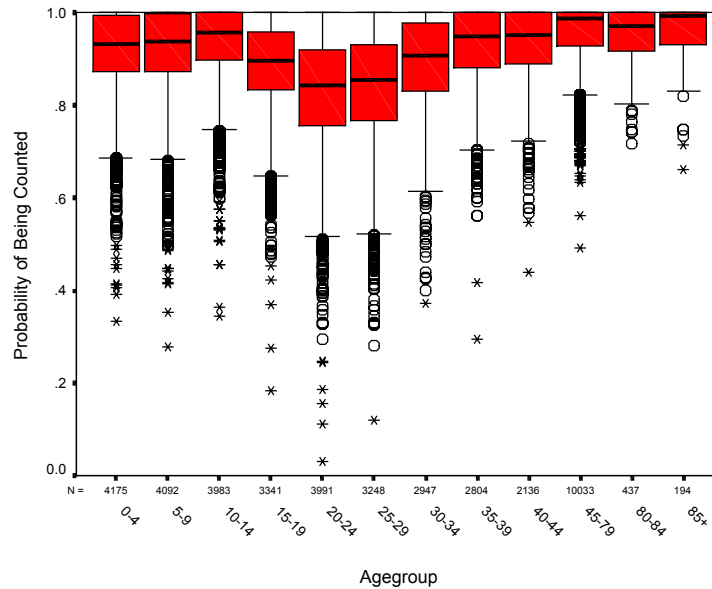


Figure 3. Distribution of probabilities of being counted for males in the ‘hardest to count’ enumeration districts.

Figure 3 shows that even after controlling for age, sex, and HtC index the individual probabilities still vary quite considerably. Some of this will be due to the primary activity last week variable. However, there is enough variability so that after controlling for this as well the probabilities will still vary which reflects what is known about the real world. This is a good property as it prevents the simulation giving over optimistic results due to there being no unobserved heterogeneity for the models to deal with. Figure 3 also shows the distribution of underenumeration by age for males with a drop in the probability of young men being counted.

5.1 CCS Design and Estimation Simulation

To create a census, an independent Bernoulli trial was carried-out for each individual. Certain rules were then applied to ensure that counted households had a sensible structure. All members of a households were underenumerated 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 being underenumerated, given that other members of their household were missed, would increase.

For the CCS, the design implemented in Section 3.4 was followed. The final design and allocation was given in Table 2. The design in Table 2 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. The average overall coverage of the censuses was 89.9 per cent. This is quite low for an overall coverage of a local authority district. Average census coverage by age and sex varied from 85.9 per cent for males 20-24 years of age to 93.8 per cent for females 45-79 years of age. The results for the RSEs calculated from the CCSs are presented in Table 4.

TABLE 4
Mean Relative Standard Errors for 1000 simulated CCSs

Males				Females			
Age Group	Number of CCSs	Design RSE	Average ^b Estimated RSE	Age Group	Number of CCSs	Design RSE	Average ^b Estimated RSE
0-4	1000	2.73	3.35 (0.986)	0-4	1000	2.66	3.28 (0.951)
5-9	1000	3.86	3.81 (1.129)	5-9	1000	3.92	3.84 (1.505)
10-14	1000	4.52	3.67 (1.066)	10-14	1000	4.45	3.86 (1.329)
15-19	1000	4.45	3.30 (0.893)	15-19	1000	4.19	3.11 (0.975)
20-24	1000	3.33	3.05 (0.699)	20-24	1000	3.22	2.33 (0.623)
25-25	1000	3.02	2.86 (0.606)	25-25	1000	2.99	2.48 (0.547)
30-34	1000	2.92	2.89 (0.645)	30-34	1000	3.12	3.00 (0.711)
35-39	1000	3.94	3.03 (0.699)	35-39	1000	4.04	3.05 (0.764)
40-44	1000	4.18	2.66 (0.649)	40-44	1000	4.53	2.54 (0.707)
45-79	1000	2.83	1.64 (0.407)	45-79	1000	2.77	1.54 (0.384)
80-84	997 ^a	7.67	3.69 (1.295)	80-84	1000	6.24	2.53 (0.803)
85+	929 ^a	10.43	5.05 (1.828)	85+	1000	3.33	3.15 (0.869)

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.

Table 4 shows that the procedure does well on average, even though the census coverage is quite low and there is additional heterogeneity (due to other characteristics) not accounted for by the model. This is because in nearly 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 the 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_c . 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. This is due to high conditional variances caused by heterogeneity amongst the probabilities of being counted. In those instances applying the stratified expansion estimator may be more efficient. Alternatively more strata could be formed before applying the regression estimator or equivalently include more explanatory variables in (2).

The simulation shows that for a perfect CCS the proposed design in conjunction with the regression estimator performs well. The more realistic situation of dependence between the census and CCS with CCS non-response has been examined in detail for the regression estimator model but is not presented here. This work has included the use of Dual System and Triple System estimation techniques. (See the series of ONC Working Papers

available from ONS for details of estimation when CCS non-response is correlated with census underenumeration.)

5.2 Small Area Simulation

The same underlying method used for the CCS design simulation in Section 5.1 was used here. Each individual in the true population had the same probability of being counted in the census. For this simulation 10 censuses, each with its own CCS, were simulated. This was less than in the CCS design simulation due to the need to keep the individual level data which is computationally much more time consuming than the totals needed for the local authority district 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 on the small area models. The simulation presented here is used to demonstrate that in general the proposed model is sensible and to investigate whether the addition of random effects improves the accuracy of the adjusted counts.

For each census-CCS pair, a matching procedure was carried out to determine the response state (8) to which each individual belonged. Initially, the fixed effects version of the multinomial model (9) was fitted to each pair. The explanatory variables used were age group, sex, HtC index, and primary activity last week. This was the same HtC index as that described in Section 3. Two adjustments were necessary: the first was that children aged 5-15 were constrained by the simulation only to be underenumerated when their household was missed. Therefore they only have two response categories: counted and missed household. The second was that all children under 16 were allocated the same category for primary activity last week. To address these two problems three models were fitted. The first model was a multinomial model for children aged 0-4. (Sex was not found to be significant and therefore it was dropped.) The explanatory variables were HtC index and sex. The second was a logistic regression at the household level for children aged 5-15. The explanatory variables were number of 5-15 year olds in the household and HtC index. The third was a multinomial model for everyone else with all the explanatory variables and additionally interaction terms for males in the age groups 20-24, 25-29, and 30-34.

Once each multinomial model was fitted for a particular census the predicted probabilities were calculated using equation (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 district since $1 - \hat{\pi}_{0ijklm}$ is the probability of being missed in the census. For the logistic model equation (10) with $r = 1$ was used to calculate $\hat{\pi}_{0jklm}$, the probability of household j being counted. The number of households was then adjusted according to size and HtC index. Using the adjusted number of households along with their size allowed for the calculation of an individual adjustment weight, which was then applied uniformly across age and sex to make the individual adjustments.

After applying the adjustment procedures this gave enumeration district counts adjusted by age, sex, HtC index, and Primary Activity Last Week for each census. Initial analysis showed very little variation between censuses. Root Mean Squared Error (RMSE) was used to assess the overall performance of the adjustment procedure. This is calculated as:

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

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

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

Again, the observed count can be either the adjusted count or the unadjusted census count. When using the unadjusted census count it should be remembered that the simulation forces each census to have a negative bias due to the fact that people are missed but no over-enumeration 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. Both (15) and (16) can be reported on the relative scale by dividing by the average true value. This can be useful when comparing across groups that vary dramatically in absolute size.

The whole model fitting and adjustment process was then repeated allowing for a random effect at the enumeration district level. The random effects models were fitted using a computer package called MLn. A discussion of MLn's estimation procedures can be found in Goldstein (1995). For each model the enumeration district residuals were estimated as well as the fixed parameters. Calculation of the necessary probabilities was straightforward for the sampled enumeration districts by applying (10) or the logistic regression equivalent. For the non-sampled enumeration districts the residual was estimated by averaging the four closest sample enumeration district residuals.

5.3 Discussion of Results for the Fixed Effects Models

The following discussion presents results for the overall adjustment described in Section 4.2, not adjustments split by counted and missed households. Its role is to demonstrate that the concept works. The above measures have been calculated for adjusted counts using the twenty fitted versions of the fixed multinomial model (9) and the ten logistic models resulting from the simulation. These have been done by age, sex, and HtC index as well as by primary activity last week, sex, and HtC index. A selection of the graphs is presented to demonstrate the results. In Figure 4 the bias for those aged 45-79 has been divided by seven to account for the much larger age group.

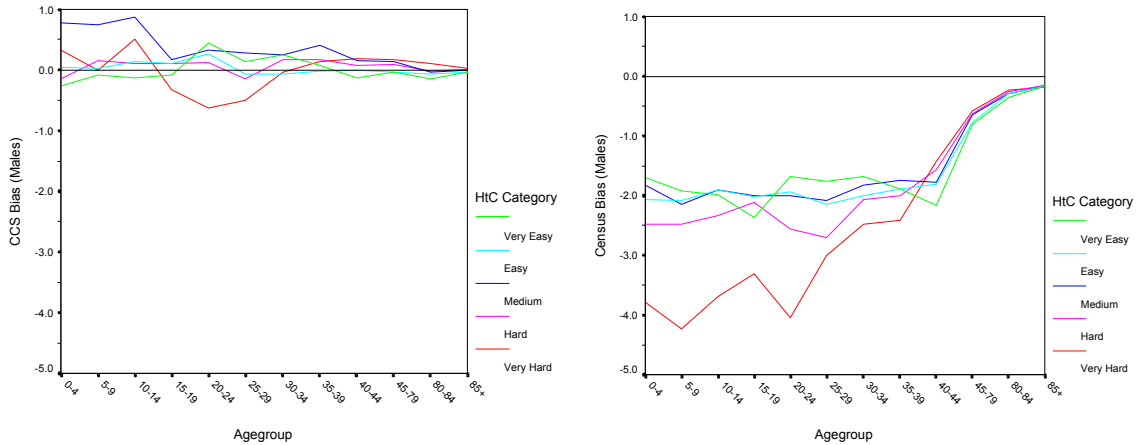


Figure 4. Bias of enumeration district totals for males by age group.

Figure 4 shows that in terms of bias, the adjusted counts for age by HtC index for males are never worse than the unadjusted counts and usually better. For young males, particularly those in the hardest to count areas, Figure 4 clearly shows the added value of the adjustments. The bias graphs for females are similar to this in terms of shape although the actual absolute value of the bias is smaller. In terms of RMSE there is little change between the adjusted and unadjusted counts. However, from Figure 4 it can be seen that its composition is changing with the reduction in bias after adjustment.

Figure 5 gives the relative bias for primary activity last week by HtC index for females. The relative scale has been used as the true counts vary considerably across the categories of primary activity last week. Figure 5 again clearly shows the gain of adjustment by reducing the relative bias in nearly all cases. This is particularly true for those activities given an underenumeration factor less than one in Table 3. The exception is the ‘other’ category. This is a very small group and for the modeling this group was collapsed with the reference category, working full-time. Figure 5 suggests that this is not appropriate for all the HtC categories.

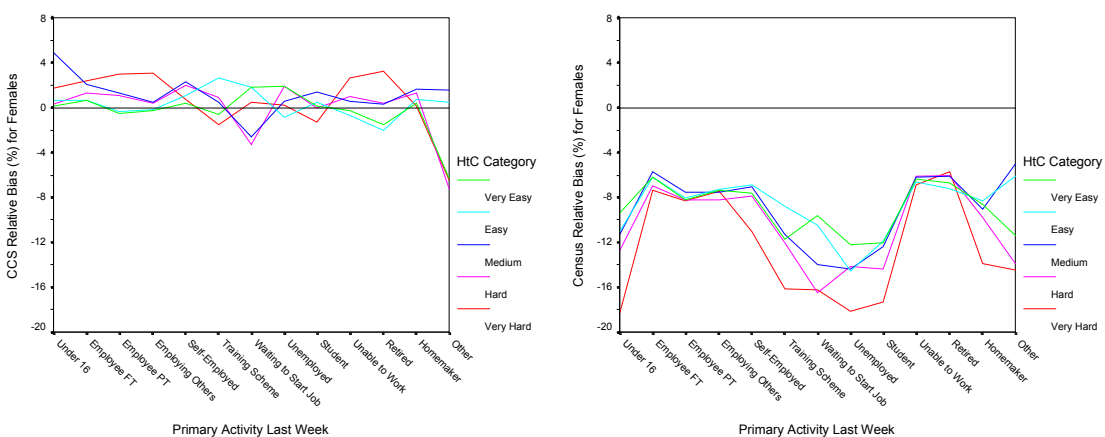


Figure 5. Relative bias of enumeration district totals for females by primary activity last week.

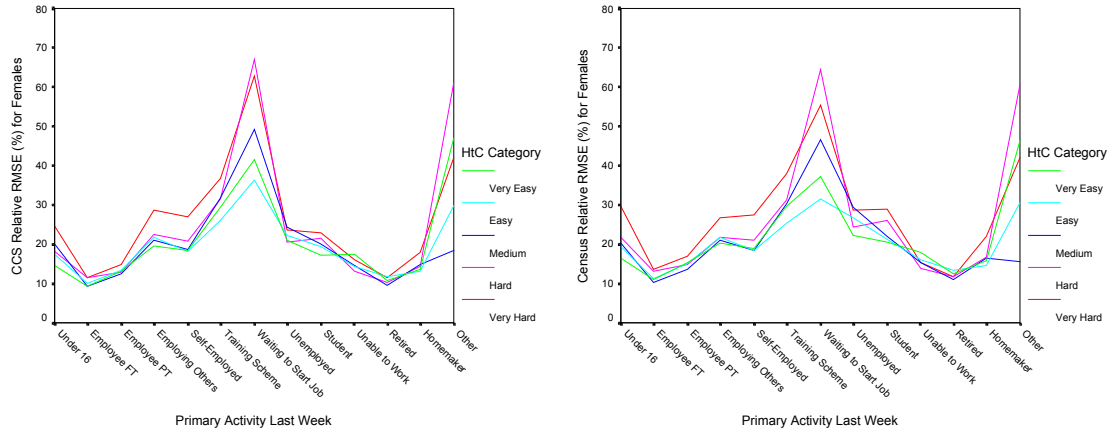


Figure 6. Relative RMSE of enumeration district totals for females by primary activity last week.

Figure 6 demonstrates that there is little change in terms of RMSE when adjusting. In fact for the waiting to start job category it is slightly worse. This is probably due to the small numbers in this particular category. The small numbers meant that for several models one or both of the multinomial model parameters was estimated as minus infinity due to a sampling zero. In these cases the particular sample suggests a zero adjustment which will not be appropriate for the rest of the enumeration districts. In addition the small sample sizes make the model parameters more variable. This same pattern is observed for the adjusted counts for males.

Figures 4 to 6 show that for the marginal distributions of age and primary activity last week the adjustment process gives good improvement in terms of bias. Table 5, at the end of this paper, gives the bias for the joint distribution of the two variables for males from the hardest to count enumeration districts. Table 5 shows that in general the bias for the adjusted count (in black) is an improvement in absolute terms on the bias for the unadjusted count (in red). The highlighted cells are those situations where this is not true. However, these highlighted cells are peculiar categories such as retired 16-19 year olds and unemployed 79-84 year olds. In principle these are possible outcomes and they do occur on rare occasions but the adjustment model will not get these peculiar cells without fitting special interaction terms. As individuals with these characteristics are rare fitting the model with structural zeros for the rare cells may be a better solution. This allows for people to have these rare sets of characteristics in the census but does not attempt to adjust the cell counts.

5.4 The addition of random effects

The results presented in this section are at an early stage. The aim is to assess whether there is an advantage to using the random effects modeling given that the adjustments based on ordinary fixed effect modeling presented in the previous section give good bias reduction. To get a feel for any gain from the addition of random effects the bias and RMSE were calculated for the enumeration district totals across HtC index for the unadjusted counts and the two adjusted counts. Initial work showed that in a few cases the adjusted counts from the spatial models were extremely high. In the case of the random effects adjustment 39 out of the 9300 enumeration districts (930 for each census) were

excluded, as their adjusted count was more than twice the unadjusted count. The relative bias is given in Table 6 with the relative RMSE in Table 7.

TABLE 6
Relative bias for enumeration district totals

Adjustment Procedure	HtC Index				
	Very Easy	Easy	Medium	Hard	Very Hard
Fixed	0.03	0.09	2.19	0.95	1.07
Random	1.57	0.39	1.94	0.31	-1.37
Unadjusted	-8.42	-8.87	-8.69	-9.99	-14.10

TABLE 7
Relative RMSE for enumeration district totals

Adjustment Procedure	HtC Index				
	Very Easy	Easy	Medium	Hard	Very Hard
Fixed	11.31	12.12	12.24	12.41	15.33
Random	13.81	13.88	13.88	12.90	15.21
Unadjusted	13.49	14.08	13.62	15.03	20.05

The results in Table 6 are disappointing as they show the adjustment based on the random effects in general not being consistently better than the fixed effects adjustment, although it is still significantly better than the unadjusted counts. In addition Table 7 shows that the RRMSE error is greater for the random effects adjustment than for the fixed effects adjustment.

Any gain in terms of bias reduction from the random effects adjustments must be offset against the fact that 39 outlying enumeration districts have been excluded. Table 8 gives the distribution of these enumeration districts by their HtC category and the simulated census.

TABLE 8
Distribution of enumeration districts excluded due to extreme adjustments

Census	HtC Category					Total
	Very Easy	Easy	Medium	Hard	Very Hard	
1	1	1	1	1	2	6
2	0	0	1	1	0	2
3	1	0	0	1	0	2
4	0	0	1	2	0	3
5	1	1	1	2	1	6
6	1	1	2	1	1	6
7	1	1	0	1	0	3
8	0	0	0	1	1	2
9	0	1	0	2	0	3
10	1	1	2	2	0	6
Total	6	6	8	14	5	39

Table 8 shows that at least one enumeration district from HtC category ‘hard’ is excluded every census. In eight of the censuses this is enumeration district ZBGJ23. Looking at this particular enumeration district the adjustment is extremely large for census six. Table 9 presents the model parameters needed to calculate the probability of an employed female aged 45-79 being counted from enumeration district ZBGJ23 by census six.

TABLE 9
*Comparison of a sample of model parameters
for enumeration district ZBGJ23 in census six*

	Parameter	Type of Adjustment	
		Fixed	Random
Model 1	Constant	-3.53	-3.73
	HtC Category 4	-0.30	-0.32
	Random Effect	0.00	5.54
Model 2	Constant	-3.97	-4.13
	HtC Category 4	-0.00	-0.07
	Random Effect	0.00	7.82
	Probability of being counted	0.96	0.02

Table 9 shows that while the model parameters for the fixed part are not identical it is the random effect that has been estimated by MLn (this is a sampled enumeration district) that really makes the difference. For this enumeration district the observed probability of being counted is low, 0.56 for the group in Table 9, suggesting that this enumeration district was given a large random effect when the probabilities of being counted were assigned in Section 5. While MLn is clearly trying to account for this the random effect assigned is too extreme. From the fixed effect model the adjustment is too small but the negative bias this introduces is not as significant as the positive bias introduced by the random effects model.

The problem of MLn predicting outlying random effects can be seen in Figure 7.

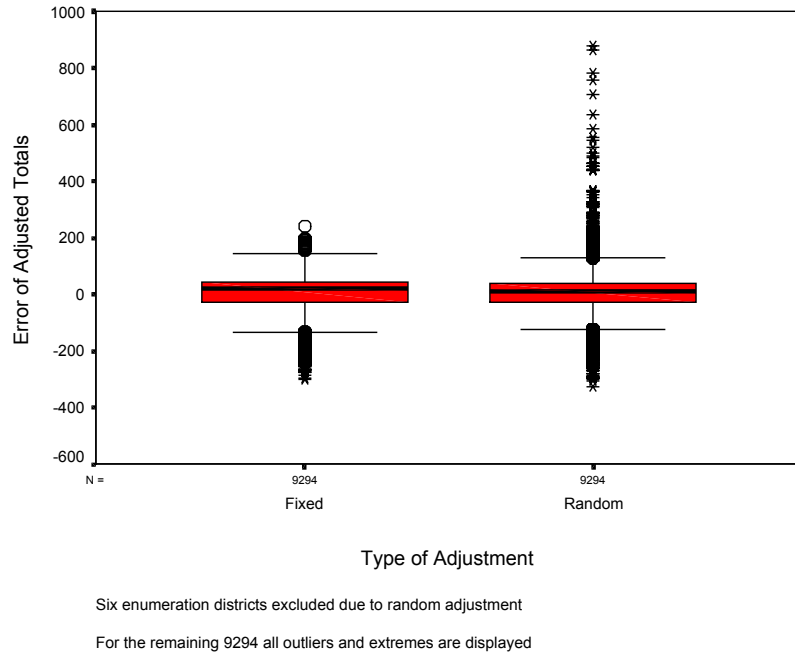


Figure 7. Distribution of errors between the adjusted enumeration district totals and the true enumeration district totals for the two types of adjustment.

Figure 7 clearly shows that the distribution of the error for the random effects adjustment procedure is highly skewed, even after excluding the six enumeration districts with errors over 1000. However, it is possible to just look at the main part of the distribution of the errors on the enumeration district totals. This is presented in Figure 8, which excludes enumeration districts that are more than 1.5 times the inter-quartile range from the upper or lower quartile.

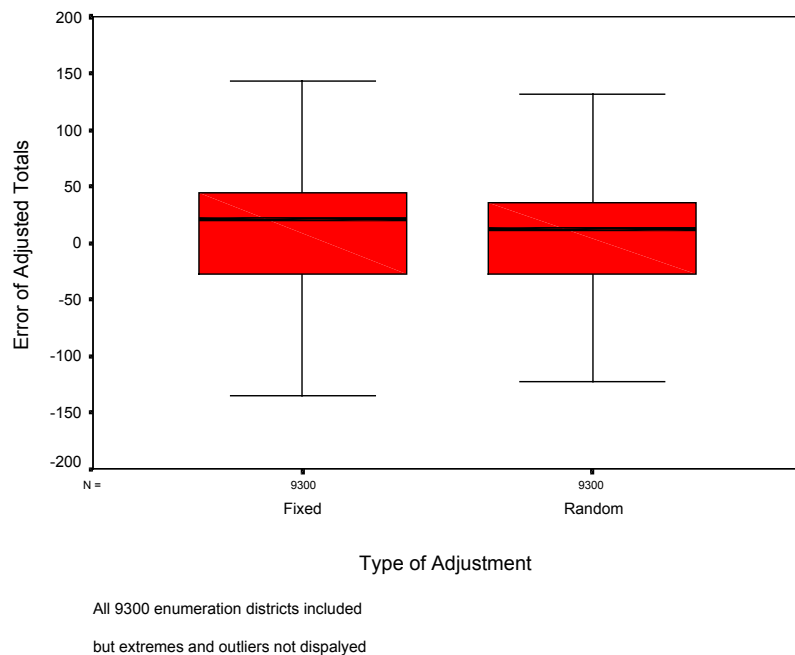


Figure 8. Distribution of errors between the adjusted enumeration district totals and the true enumeration district totals for the two types of adjustment with outliers and extremes not displayed.

From Figure 8 it is possible to see that in the majority of cases the random effects adjustment does do better. The median error is closer to zero, 12.5 for the random effects and 20.6 for the fixed effects, and the inter-quartile range represented by the box in the plot is smaller. Therefore, for the majority of cases the random effects adjustment is more efficient. However, it is sensitive to outliers. In general the fixed effects adjustment doesn't fit as well for individual enumeration districts but it is more robust to the extreme cases.

6. DISCUSSION AND CONCLUSIONS

This paper has described a portion of the research so far undertaken to develop a strategy for a 'One Number Census' in England and Wales in 2001. The major tool of this strategy will be a post enumeration survey, to be called the Census Coverage Survey, which will constitute a radical change from previous censuses. It will focus purely on coverage and will comprise a short questionnaire addressing the characteristics most associated with underenumeration. The survey will aim to make estimates, by age and sex, of underenumeration for sub-national areas. The paper has described the design of this survey and the simulation study in Section 5.1 shows that for a perfect CCS response rate accurate sub-national population estimates, adjusted for underenumeration can be obtained. This work has been extended to include CCS non-response although this has not been presented in this paper.

Given agreed sub-national estimates of underenumeration it will then be necessary to make estimates at an individual level. The paper has focused on fixed and multilevel regression approaches to estimate an individual's probability of being counted in the census. The results of the simulation study in Section 5.3 to assess the potential of this regression approach are very promising. For the fixed effects model they show that it is possible to adjust for missing people at the enumeration district level and reduce the bias while not increasing the total error, as measured by RMSE. Applying these probabilities to fully adjust the census database through imputation or weighting is an area that still requires extensive research.

The addition of random effects, spatially smoothed to give estimates of the residuals for non-sampled postcodes, was expected to improve the results as the random effect allows for additional unexplained small area variability. In reality there has not been an overall improvement to the adjusted counts. Investigation has attributed this to MLn being outlier sensitive when the estimated random effects are used for prediction. However, for the majority of enumeration districts the addition of random effects has reduced the error on the adjusted enumeration district totals. Further work is now needed to investigate a more robust method for estimating the random effects for sampled enumeration districts.

One suggestion that is being considered is to constrain the fixed parameters in MLn to be the same as those for the fixed effects adjustment. Therefore, MLn would just be estimating the random effect. However, this ignores the fact that the way MLn is estimating this enumeration district random effect is different to the way it was introduced

into the probability of an individual being counted in Section 5. Considering this fact a simpler approach is also being considered which will use the average of the errors from the fixed effects model to estimate an enumeration district effect that can then be smoothed. It is expected that this simpler approach will give some of the advantage of the random effects procedure for the majority of enumeration districts while being robust to outliers.

7. REFERENCES

- Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L. and Zaslavsky, A. M. (1993) Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *J.A.S.A.*, **88**, 1149-1159.
- Charlton, J., Chappell, R. and Diamond, I. D. (1997) Demographic analyses in support of a One Number Census. *Proceedings of the Statistics Canada Symposium*.
- Goldstein, H. (1995) *Multilevel Statistical Models Second Edition*. London: Arnold.
- OPCS (1993) Rebasing the annual population estimates. *Population Trends*, **73**, 27-31.
- OPCS (1994) Undercoverage in Great Britain. *1991 Census User Guide 58*, London: HMSO.
- 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.

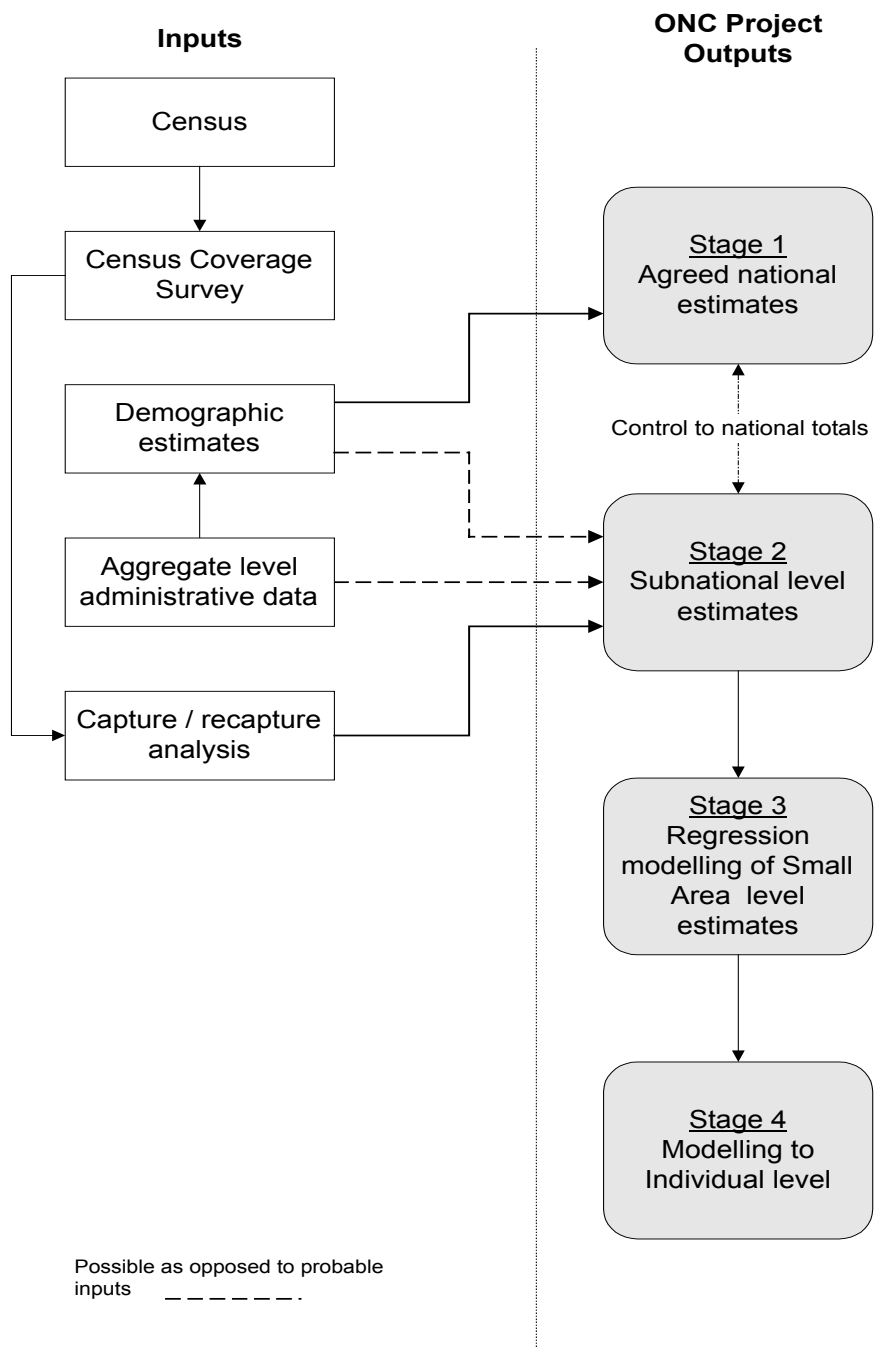


Figure 1. The stages of a One Number Census

TABLE 5

Bias by age and primary activity last week for the adjusted and unadjusted counts of males in HtC category five

Economic Status	Age Group (Years) for Males in the Hardest to Count Category of the HtC Index								
	16-19	20-24	25-29	30-34	35-39	40-44	45-79	80-84	85+
Employed (full-time)	-0.01 -0.36	-0.48 -1.71	-0.30 -1.50	-0.08 -1.12	-0.06 -1.10	0.05 -0.71	0.29 -0.87	0.07 -0.02	0.09 0.00
Employed (part-time)	-0.06 -0.20	-0.03 -0.20	-0.07 -0.23	0.05 -0.13	-0.01 -0.21	0.10 -0.09	0.06 -0.11	0.03 -0.07	0.10 0.00
Employing Others	-0.17 -0.29	-0.06 -0.20	-0.11 -0.25	0.01 -0.20	0.05 -0.18	0.00 -0.17	0.08 -0.14	0.09 0.00	0.09 0.00
Self Employed	0.04 -0.11	-0.01 -0.23	-0.06 -0.29	-0.05 -0.37	0.07 -0.29	0.03 -0.20	0.07 -0.22	-0.03 -0.12	0.10 0.00
Training Scheme	-0.04 -0.28	-0.04 -0.27	-0.02 -0.25	-0.04 -0.24	0.05 -0.19	0.06 -0.13	-0.11 -0.23	- -	- -
Waiting to Start Job	0.00 -0.21	-0.06 -0.23	-0.04 -0.22	0.02 -0.23	0.16 -0.13	0.00 -0.22	-0.02 -0.16	- -	- -
Unemployed	-0.05 -0.59	-0.02 -1.06	-0.02 -0.79	0.02 -0.77	0.06 -0.73	0.09 -0.42	-0.03 -1.06	0.18 0.00	-0.78 -0.82
Student (full-time)	-0.28 -1.71	-0.08 -1.09	-0.08 -0.52	0.04 -0.36	0.02 -0.41	0.01 -0.26	-0.08 -0.27	- -	0.10 -0.05
Unable to Work	-0.04 -0.17	-0.02 -0.17	-0.03 -0.19	0.04 -0.16	0.02 -0.25	0.03 -0.18	0.05 -0.60	0.08 -0.02	-0.10 -0.18
Retired	0.14 0.00	-0.03 -0.13	0.00 -0.11	0.17 0.00	0.25 0.00	-0.04 -0.20	0.77 -1.05	0.10 -0.21	0.05 -0.13
Homemaker	0.12 -0.06	0.00 -0.18	-0.20 -0.35	0.00 -0.19	0.12 -0.15	0.09 -0.12	-0.02 -0.18	-0.05 -0.16	- -
Other	-0.04 -0.20	-0.19 -0.31	-0.07 -0.20	-0.09 -0.21	0.10 -0.05	-0.01 -0.12	-0.10 -0.19	- -	0.09 0.00