ONS(ONC(SC))97/13

ONE NUMBER CENSUS STEERING COMMITTEE

Modelling to small areas - a full One Number Census

1. The attached paper reviews the research into the modelling at small areas and the adjustment of records as part of a One Number Census. This is a new approach to that presented at the last Steering Committee meeting as Working Paper 2 (ONS(ONC(SC))97/03).

2. Further work is planned to:

   a) increase the complexity of the exclusion model for creating Censuses, which will include introducing small area spatial variation;

   b) introduce undercoverage into the Census Coverage Survey (CCS);

   c) introduce dependence between the Census and CCS as investigated in ONS(ONC(SC))97/10 and ONS(ONC(SC))97/12.

3. The Steering Committee are asked to:

   a) note the paper

   b) provide any comments (at the forthcoming meeting or in writing by 10 December 1997) on the proposed plans for further research.

Lisa Buckner
Census Division
Office for National Statistics

Room 4200W
Segensworth Road
Titchfield
Fareham
Hants PO15 5RR

November 1997

Modelling to small areas - a full One Number Census

James Brown, Lisa Buckner, Ian Diamond and Ray Chambers

1. Introduction

1.1 The ultimate aim of the ONC project is a single Census database fully adjusted for underenumeration. This requires a procedure that allows the imputation of missing people at a very small area for both counted households and missed households. Previously, the solution to this problem presented in Working Paper 2 (ONS(ONC(SC))97/03) involved the use of two logistic regression models. Initial work in setting up the simulation presented later in this paper, revealed major practical difficulties to this approach. As a result of this, a new approach was investigated and this is presented below.

2. Situation after the Census and CCS

2.2 Let us assume that the Census Coverage Survey (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** Census individuals with any **extras** from the CCS. This is a slightly different assumption to the one in paper ONS(ONC(SC))97/10 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.

2.2 At the individual level one has:

i) a matrix of socio-economic characteristics **X**
(age, sex, marital status, ethnicity, economic status)

ii) a matrix of household characteristics **Z**
(tenure, building type, multiple-occupied, number of residents)

iii) a vector of the household structure **S**

The household structure vector indicates the type of social structure between individuals within the household such as:

Single Person
Couple With No Children
Nuclear Family (Couple + Children)
Extended Family (Couple + Children + Others)
Single Parent Family
Household of Unrelated Members
Communal Establishment (Institution).

Each individual i belongs to a household j within postcode k within enumeration district l of district m. The CCS does not contain all districts or postcodes so there is a

prediction problem for the non-sampled postcodes. From the CCS direct estimation, demographic analysis, and capture recapture modelling there are **gold standard** age sex totals. The goal is to share the 'extra' people amongst the enumeration districts.

3. Multinomial model for small area adjustments

3.1 In relation to the assumption above, consider the possible categories of enumeration into which an individual can fall. A person is either counted, missed in a counted household, or missed in a missed household. This can be represented by the dependent variable Y_{ijklm} where:

$Y_{ijklm} = 0$ when individual i is counted in the Census (but not necessarily the CCS as well)

$Y_{ijklm} = 1$ when individual i is missed in the Census and household j is counted (with respect to the CCS)

$Y_{ijklm} = 2$ when individual i and household j are missed in the Census (with respect to the CCS)

This is a multinomial variable where:

$P(Y_{ijklm} = 0) = \pi_{0ijklm} = P(i \text{ is counted})$

$P(Y_{ijklm} = 1) = \pi_{1ijklm} = P(i \text{ is missed} \cap j \text{ is counted})$

$P(Y_{ijklm} = 2) = \pi_{2ijklm} = P(i \text{ is missed} \cap j \text{ is missed})$

$\pi_{0ijklm} + \pi_{1ijklm} + \pi_{2ijklm} = 1$

and in general these probabilities will depend on the characteristics of the person, household, postcode, etc. Putting aside measurement error problems[1] the following multilevel multinomial model can be fitted for the CCS sample postcodes:

[1] In general, it is of course very likely that there may be different responses for the Census and CCS respectively. Future research will develop a set of criteria which will be used to address this measurement problem. This impacts on interpretation of model parameters and prediction for non-sampled postcodes when only Census information is available.

$$\ln\left(\frac{\pi_{1ijklm}}{\pi_{0ijlkm}}\right) = \alpha_1 + \underline{\beta_1}' \underline{X}_{1ijklm} + \underline{\gamma_1}' \underline{Z}_{1jklm} + \underline{\eta_1}' \underline{S}_{1jklm} + \sigma_{1m} + \varsigma_{1lm} + \lambda_{1klm} + \upsilon_{1jklm} + \varepsilon_{1ijklm}$$

$$\ln\left(\frac{\pi_{2ijklm}}{\pi_{0ijlkm}}\right) = \alpha_2 + \underline{\beta_2}' \underline{X}_{2ijklm} + \underline{\gamma_2}' \underline{Z}_{2jklm} + \underline{\eta_2}' \underline{S}_{2jklm} + \sigma_{2m} + \varsigma_{2lm} + \lambda_{2klm} + \upsilon_{2jklm} + \varepsilon_{2ijklm}$$

This is a standard random intercepts model and is important for small area estimation as this allows for extra heterogeneity between postcodes and enumeration districts. In fitting the final model the possibility of random coefficients will, of course, be addressed.

4. Prediction for non-sampled postcodes

4.1 As not all postcodes are in the sample, the first stage is to use $\underline{\hat{\beta}_1}$, $\underline{\hat{\beta}_2}$, $\underline{\hat{\gamma}_1}$, $\underline{\hat{\gamma}_2}$, $\underline{\hat{\eta}_1}$ and $\underline{\hat{\eta}_2}$ to get predicted probabilities for each of the different types of individuals and households in all areas. Again ignoring measurement error issues, this is straightforward for the fixed effects model but not for the multilevel model. For the latter case there is no estimate of higher level residuals. This is due to the independence assumption made in the multilevel framework. Ideally one would like to fit full spatial[2] random effects. Computationally speaking this is currently extremely difficult. A proposal which is currently being considered is to fit the model in the independence framework and then use a spatial[2] smoothing function to estimate residuals for non-sampled postcodes assuming the random effects are significant. This means that in principle for all areas you can estimate $\hat{\pi}_{0ijklm}$, $\hat{\pi}_{1ijklm}$, $\hat{\pi}_{2ijlkm}$.

5. Adjusting the Census

5.1 Let N_{ijklm} be the Census count of individuals with the set of characteristics given by ijklm. (eg. white 20-24 married employed male renting a detached house who is a member of a nuclear household of size 3 within postcode k.)

P(people of type ijklm are counted) = $\hat{\pi}_{0ijklm}$

implies P(people of type ijklm are missed) = 1 - $\hat{\pi}_{0ijklm}$

From this the number of people of type ijklm who are missed is given by:

---

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

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

5.2 The problem now is how to allocate these 'extra' people to already counted households or completely missed households. Given that an individual is missed one requires the probability that their household was missed or counted.

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

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

***NOTE*:** P(j is counted | i is missed) + P(j is missed | i is missed) = 1 as required.

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

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

and the number of missed people from missed households is:

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

where $N_{ijklm} = N_{1ijklm} + N_{2ijklm}$ and the adjustments come directly from the multinomial model.

6. Locating extra people

6.1 For the people from counted households (N_{1ijklm}), the task is to search the postcode for suitable households based on the household characteristics, select a household using a random number generator or nearest fit criterion, and add the person. In certain cases the donor households will need a different structure to that of the new person. For example the donor household for a married man from a nuclear family would be a single mother in the unadjusted Census.

6.2 For the people from the missed households, there will be a set of groups of people given by the different N_{2ijklm}s. The task is then to fit the individuals back together as households. One possible way would be through a simulation which built-up households from available individuals. Another solution would be through some kind of iterative proportional fitting algorithm where the N_{2ijklm}s form marginal totals for types of individuals and the cells would be completed households.

6.3 There is a technical problem with the model proposed above. Clearly, nobody from a single household can have $Y_{ijklm} = 1$. For the model to be estimated it may be necessary to 'introduce' a small number of artificial cases but one would expect π_{1ijklm} to be very close to zero.

7. Simulation Study Methodology

7.1 The same underlying method used for the CCS design simulation was used here. Each individual in the true population had the same probability of being counted in the Census. Initially 10 Census - CCS pairs 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 first simulations presented here the Census and CCS were assumed to be independent with perfect coverage for the CCS.

7.2 For each Census CCS pair, a matching[3] procedure was carried out to determine the multinomial response category for each individual. The fixed effects version of the multinomial model described in Section 3.1 was fitted to each pair. The explanatory variables used were age group, sex, and Hard to Count (HtC) index. This was the same HtC index as that used in paper ONS(ONC(SC))97/10. At this stage the household structure was not added since it was considered better to investigate the simplest case first.

7.3 Once the model was fitted the predicted probabilities were calculated and the adjustment for *all* missed people applied to all enumeration districts. This gave adjusted age sex enumeration district counts by HtC index for each Census. Initial analysis showed very little variation between Censuses. Therefore calculations of Mean Squared Error (MSE) and bias were carried out across enumeration districts and Censuses.

8. Assessing the Simulation

8.1 To look at the overall performance of the adjustment procedure, Root Mean Squared Error (RMSE) was used. This is a good overall measure since it includes the effect of variance and bias. Taking the square root results in it being on a comparable scale to the true counts which were being estimated. The RMSE is calculated as:

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

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

[3] In the simulation accurate matching is trivial but it is recognised that in reality this may well be the most difficult task in the whole process.

in the double sum. In the formula, the observed count can either be the adjusted Census count or the unadjusted Census count. Calculating the RMSE in each case allows comparisons to see the gains of adjustment.

8.2 It is also important to look at the bias on its own. If bias is driving the RMSE then this shows that there is a systematic effect from adjustment. Altering the model may help fix this. The bias is calculated as:

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

In this formula the observed count can again be either the adjusted count or the unadjusted Census count. The relative bias was also calculated by dividing the bias by the average true count for an enumeration district. The advantage of relative measures is that large groups get a better representation. The disadvantage is that small groups often have very large relative biases, when the actual bias is so small its overall effect is negligible.

9. Discussion of Initial Results

9.1 These results are the first stage of analysing this simulation data and only present an overall adjustment, not adjustments split by counted and missed households. Their role is to demonstrate that the concept works, not give a definitive picture of how the procedure would work in a One Number Census. The above measures have been calculated from the 10 models resulting from the simulation. The results are presented separately for males and females to allow comparisons. The RMSEs are shown first as these are considered the best overall measure of performance.

Figure 1

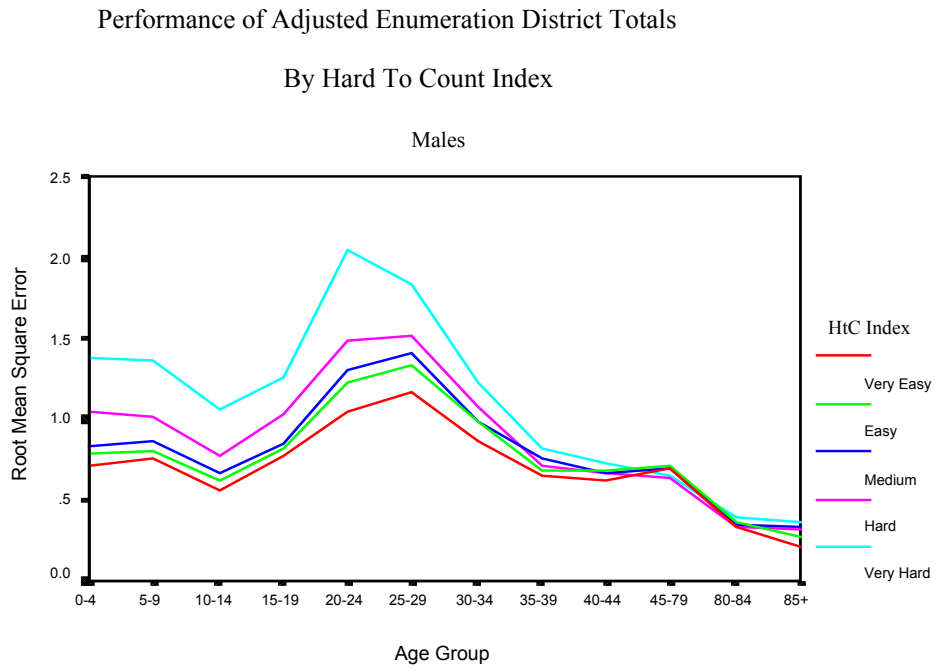
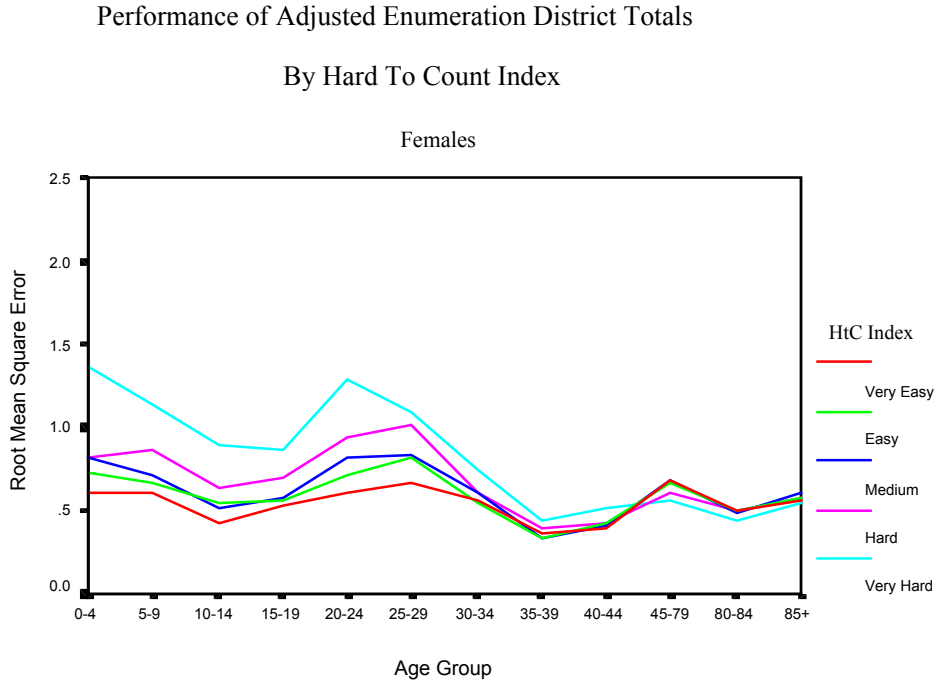


Figure 2



Figures 1 and 2 for RMSE show in general females doing better than males with the key ages of 20-34 being the worst estimated. This reflects the fact that within these age groups the most people, in absolute numbers, need to be estimated. 0-4 year olds are the same for both but comparatively they are more important for females as females have a lower undercount than males, in absolute numbers, for the other ages. The index groups behave as expected with the hardest to count having the highest RMSE.

From Figures 1 and 2 it can be seen that even for the hardest to count group males for the RMSE stays between 1 and 2. This represents an error of between 1 and 2 people on average across the enumeration districts as a result of variability and bias. For other groups it is below 1. It is of use to see the bias on its own as this can seriously effect confidence interval coverage if it dominates the variance. In reality it is non-trivial, if not impossible, to estimate the bias as the true value is not known. Figures 3 and 4 present the bias for males and females.

Figure 3

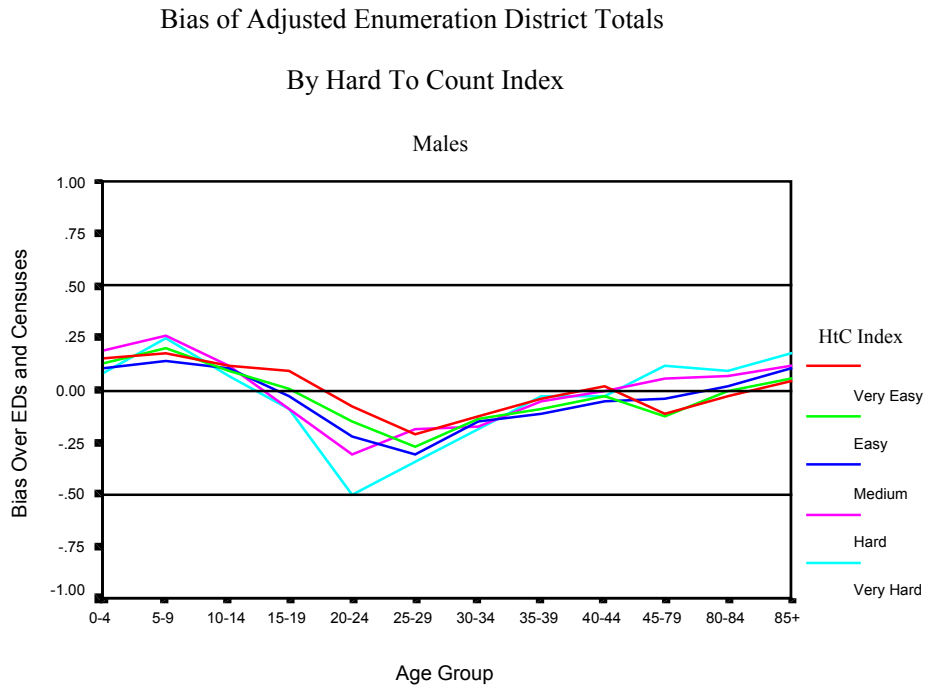
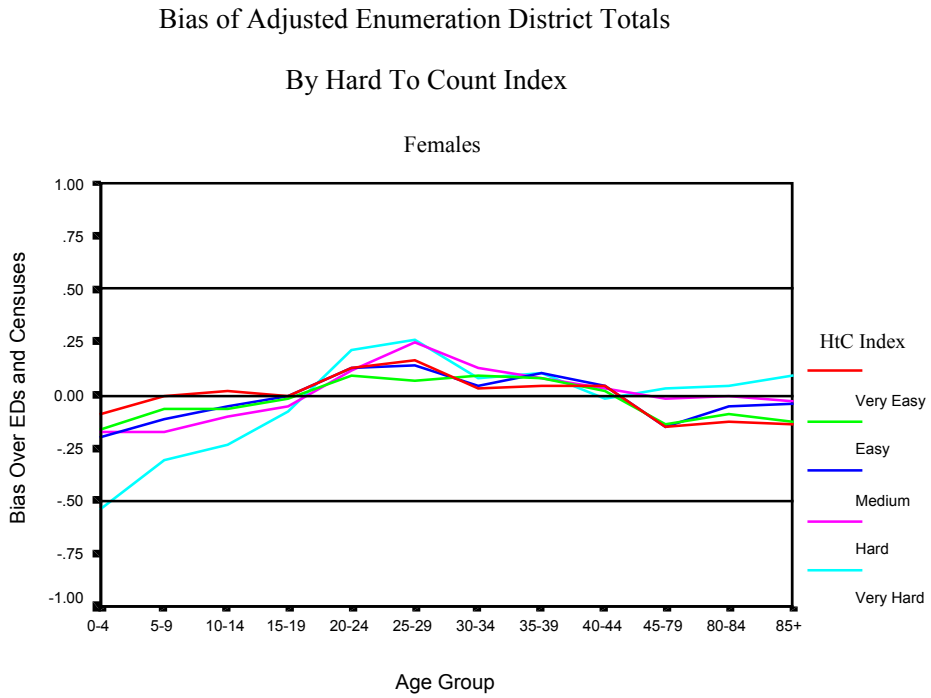


Figure 4



9.2 The bias shows a slightly different picture. As expected 20-34 year old males are being systematically underestimated but the comparing the bias to the RMSE in figure 1 shows that it is not dominating the overall performance. What is less expected is the strong negative bias for 0-4 year old females compared to the only slight positive bias for males of the same age, when the RMSE in each case is similar. This shows that the bias for the females is having a comparatively larger effect on the RMSE than it is for the males.

From the absolute size of the bias, the worst case is that on average ½ a 0-4 year old female and ½ a 20-24 year old male are missed for each very hard to count enumeration district. Showing the relative bias puts these and the other biases into context as a proportion of the total number of people you are trying to estimate. The relative biases are presented in Figures 5 and 6.

9.3 Figures 5 and 6 show that for most age groups the relative bias is less than two percent of the total. It rises for the oldest age groups as the denominator is getting much smaller and missing one person out of five has a high relative importance compared to missing one out of 50. The interesting pattern for females in the 85+ group is discussed in more detail later.

Figure 5

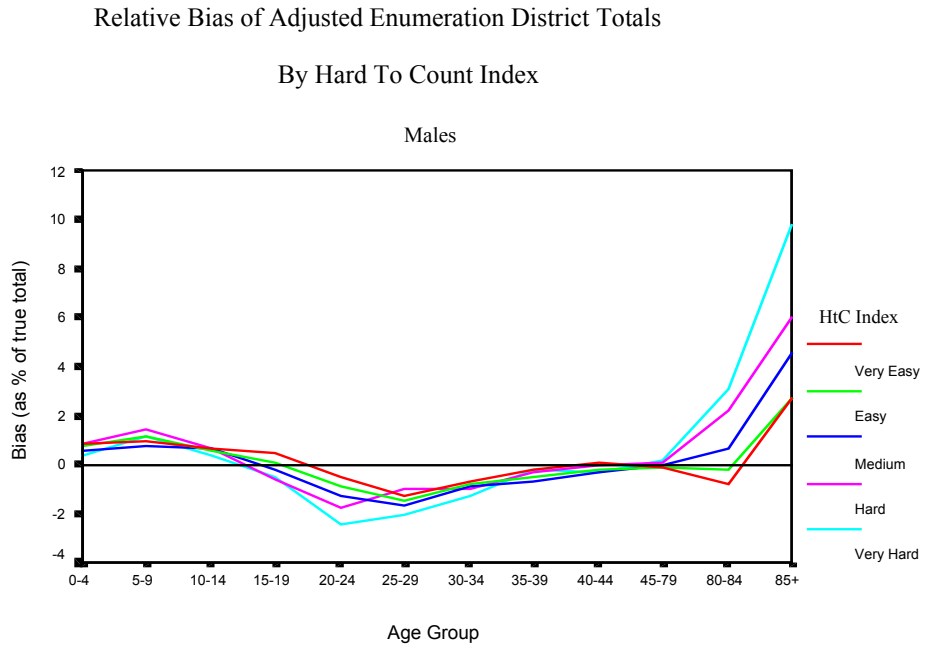
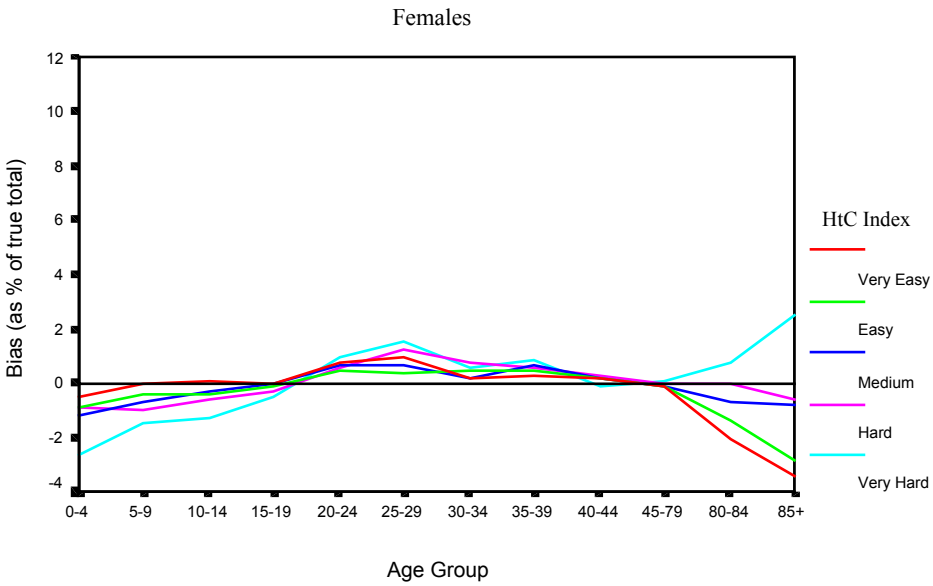


Figure 6



9.4 Figures 1-6 look only at the adjusted counts. While one can say what the RMSE means or what the bias means it is hard to say how good is this. The key question is what has been gained from doing the adjustments? To answer this question Figures 7-10 compare the adjusted counts for the hardest to count enumeration districts to their unadjusted census counts.

Figure 7

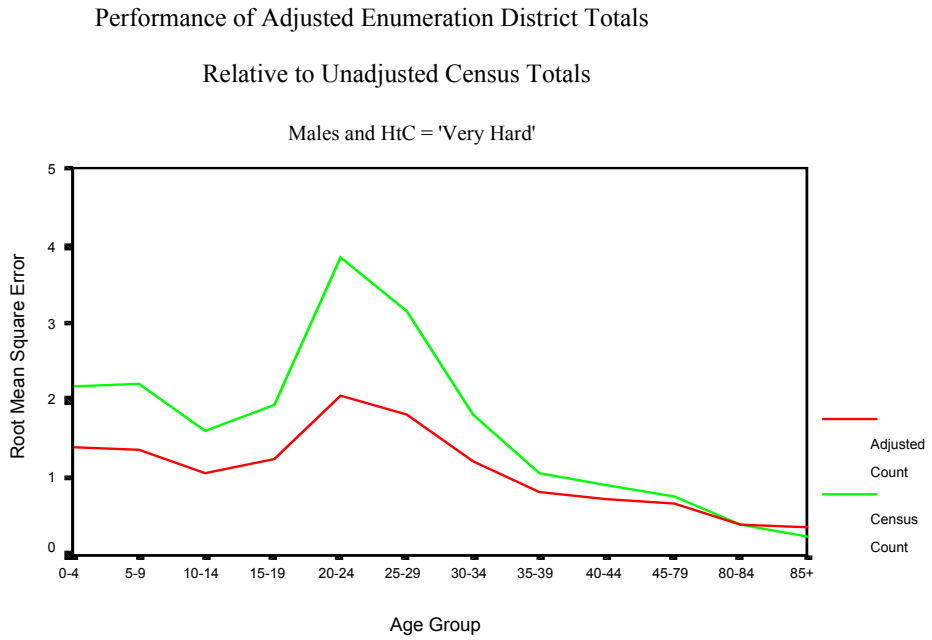
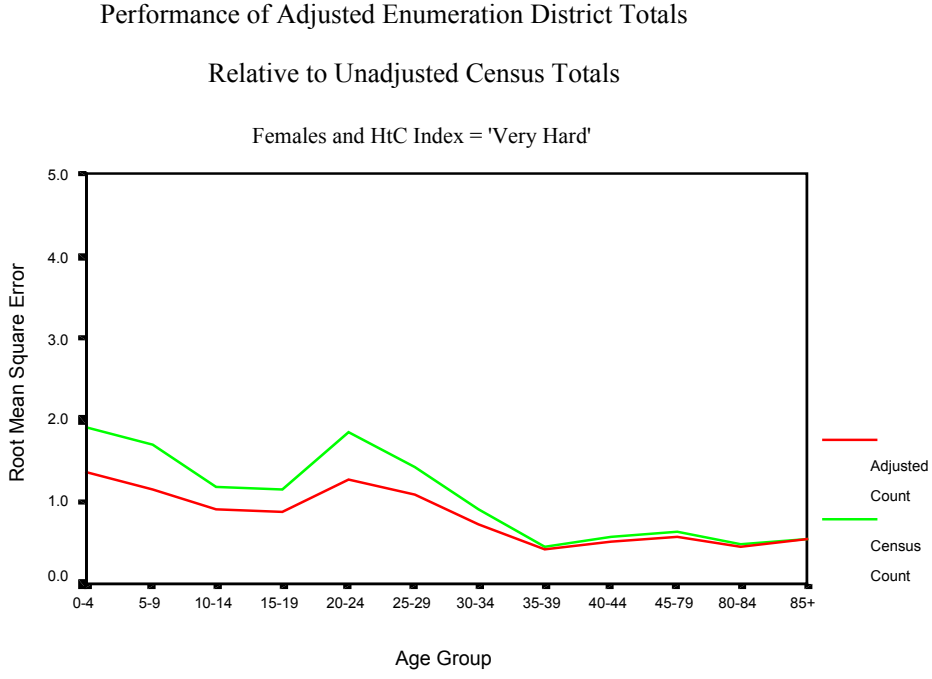


Figure 8



9.5 Figures 7 and 8 show that in terms of RMSE, the adjustment process is never worse than the unadjusted counts and usually better. This is good news and for young males Figure 7 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. Looking at the bias it can be seen that for this age-sex group the census approaches zero while the adjustment puts too many people in. In general, in terms of bias the adjusted counts are also better. This statement should be made with caution since 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 counts, the bias is averaged over positive and negative quantities.

Figure 9

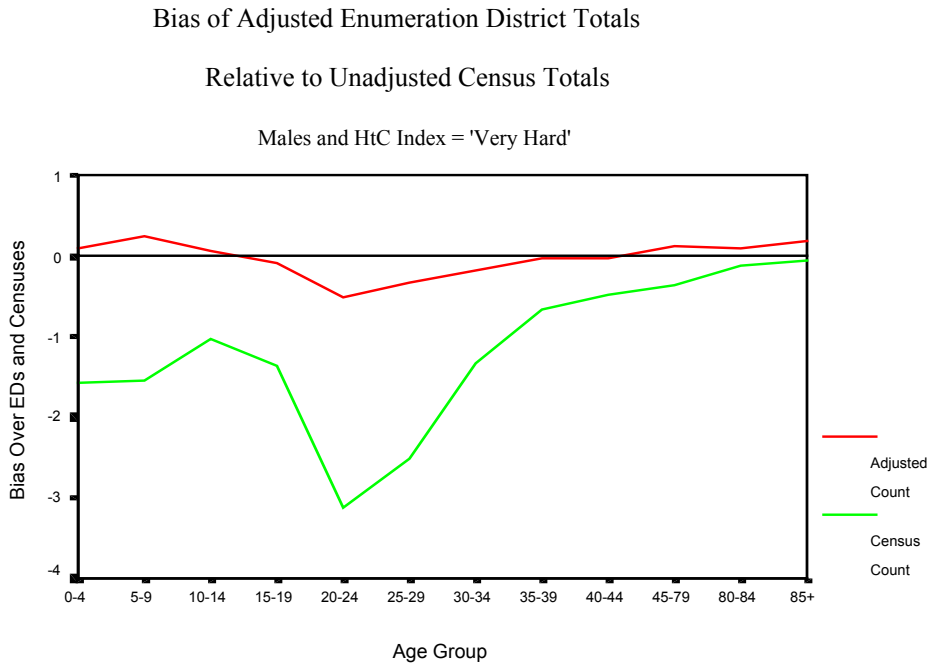
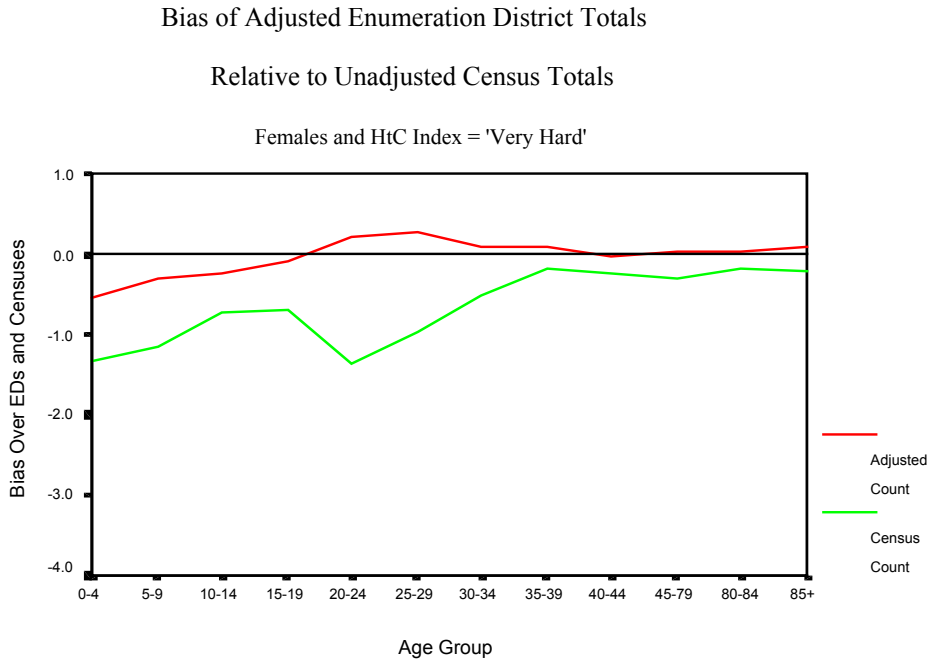


Figure 10



9.6 A possible explanation for the rather strange results for the female bias may lie in the chosen model. There are no interactions in the model. For One Number Census one would do a thorough model fit and check for interactions but at this stage it is of interest to see how well the simple model works. The main effects tend to give higher adjustments to males, higher adjustments as the index moves to harder to count

enumeration districts, and higher adjustments to 20-34 year olds. It will also slightly raise adjustments for 0-4 year olds and 85+. As there are only fixed effects, the model will give too much to young males (the sex effect) and not enough to young females. For the oldest age group, males are relatively unimportant as there are so few males in the 85+ group. The females dominate and push the age adjustments up; adding the sex and index effects leads to the positive bias for males. (This is seen best on the relative bias charts (Figures 5 and 6) where the small numbers in the denominator make the small biases relatively more different.) However, while the age adjustment for females might push the adjustments up, the sex effect pushes it down. This leads to negative bias for most index groups, except for the very hard to count group where the index effect is too strong.

9.7 To investigate this further, a one-way ANOVA was performed on the bias using age groups, sex, and index as the factor. It was also carried out for age sex groups combined.

Table 1 - ANOVA for the Adjusted Count Bias

Factor		Sum of Squares	Degrees of Freedom	Mean Square	F- Ratio	Sig.
Age	Between Groups	0.120	11	0.0109	0.486	No*
	Within Groups	2.421	108	0.0224		
	TOTAL	2.541	119			
Index	Between Groups	0.0226	4	0.00565	0.258	No*
	Within Groups	2.518	115	0.0219		
	TOTAL	2.541	119			
Sex	Between Groups	0.000805	1	0.000805	0.037	No*
	Within Groups	2.540	118	0.0215		
	TOTAL	2.541	119			
Age and Sex	Between Groups	1.952	23	0.0848	13.828	Yes**
	Within Groups	0.589	96	0.00614		
	TOTAL	2.541	119			

* 10 percent level of significance.
** 0.1 percent level of significance.

Table 1 shows that the variation in the bias is not being driven by any of the main effects that are in the model. However, the combined age sex groups are a significant

factor in a one-way ANOVA suggesting that an age-sex interaction would improve the model and reduce the variation in the bias for certain specific groups.

10. Conclusions and Further Work

10.1 These initial results are promising and show that the method works. The model is a simple fixed main effects model. The one-way ANOVA results combined with the shape of the female bias suggest that interacting sex with certain age groups (0-4, 20-34, 85+) will improve results for both males and females. In the reality of the One Number Census one would also expect to do even better by fitting random effects to account for additional small area variability. (Random effects have not been fitted yet as the current simulation has no small area variability beyond enumeration districts belonging to the same hard to count group.) These results are obtained by using $1/\pi_0$ to adjust for all missing people. The next stage is to do a similar analysis for the π_1/π_0 and π_2/π_0 adjustments for this simple model using the standard simulation.

10.2 The models were fitted using SAS which handles the situation where certain response groups do not exist and gives a warning. SAS effectively sets the value of the parameters to negative infinity, which results in predicted probabilities of approximately zero. Once models are required with random parameters it is necessary to use MLn (a multilevel modelling computer package) or some similar package. The same result can still be achieved by using the offset command to set parameters and thus stop MLn from trying to fit them.

10.3 To investigate more complex models requires a more complex simulation. The current simulation only excludes people from the Census based on age, sex and HtC index, hence these are the only variables the multinomial model uses. The next two major steps are to include household structure into the model and then extend the simulation model. Extending the simulation model will include using other variables to exclude people such as economic status, ethnicity and tenure. It will also involve introducing spatial small area effects into the data. This will allow us to investigate how strong these small area effects need to be for the fixed effects estimation model not to be sufficient and need random effects. It will also allow investigation into the value of the spatial smoothing of random effects that has been proposed. There is also the need to do sensitivity analysis of the models to dependence between the Census and CCS as well as CCS undercoverage.

10.4 The simulation shows that when people go missing by certain characteristics the multinomial model is able to recover them by modelling those characteristics. Provided the CCS collects the key variables that determine individual undercount the work so far confirms that the multinomial model is able to adjust the census for undercoverage. All this will only lead to the obtaining adjustments. There is still the question of locating individuals in counted and created households and this will form a major part of the next stage. There will be synergy here with the work of the Census Imputation team.