Estimating and Correcting for Over-count in the 2011 Census

Amy Large (ONS) and James Brown (UoS)

1) Introduction

Several papers have already discussed the issue of estimating the level of over-count in the 2011 Census (Abbott and Brown (2007), Brown *et al* (2009) and Abbott and Large (2009)). However, this paper brings much of this information together and includes a small simulation study to demonstrate the performance of the estimator proposed by Brown *et al* (2009) and how the estimate of duplication arising from the methodology described by Abbott and Large (2009) can be integrated into the overall estimate.

2) Types of Over-count

Historically, over-count has not been considered an issue in UK Censuses. Estimates for 2001 range from 0.1% (the original estimate based on the CCS) to a minimum of 0.4% (duplicates estimated from the longitudinal study) (Abbott and Brown, 2007). However, it is possible, with changes in data collection methodology, and in social behaviour, that it will be more prevalent for the 2011 Census. Four types of over-count have been identified:

Type 1: Duplicate returns at the same location.
This is where two or more returns are made in the same small area (postcode) by the same household (or some members of a household). For example, a household may return a paper form, and submit an internet questionnaire. There are processing procedures in place for 2011 in order to rectify and resolve this type of over-count as much as possible.

Type 2: Duplicate returns from different locations.
This is where a response is received for the same person but from a different small area, such as students being counted at both their term-time address (correctly) and their parents' address (incorrectly).

Type 3: Counted in the wrong location.
This is a return where the Census only counts the 'wrong' half of a duplicate. This might happen by only counting the student at their parents' address or by counting a mover after Census night only at the new address.

Type 4: Erroneous returns.
This is a return that is purely fictitious and should not appear in the Census at all at any location. This can be a joke return, a creation of the processing, a baby born just after Census day or individuals that died just prior to Census night.

Type 4 over-count is only identifiable by further field work, and cannot be accounted for through matching, estimation or adjustment.

As mentioned previously, type 1 over-count should be rectified during data processing. However types 2, 3 and 4 will remain in the Census population.
In this paper we present the main estimation framework with a simulation study demonstrating how we can a) estimate this over-count and b) adjust the estimate being produced accordingly.

3) Estimation Framework for the Level of Over-count

We expect to undertake two major matching exercises. Match one will sample Census returns and search for duplicates[1] (Abbott and Large, 2009), while match two will search the Census database for alternative returns for those sampled in the CCS[2] (Brown *et al*, 2009). This will provide information to estimate the level of over-count defined as

$$\gamma_{ag} = \frac{\text{Census count for group a in area g}}{\text{number of unique individuals for group a in area g correctly counted by the Census}}.$$

To operationalise this we define three quantities:

- $X_g^{(a)}$ is the total census count for group a in area g[3],
- $Y_g^{(a)}$ is the *correct* census count for group a in area g,
- $E_g^{(a)}$ is the erroneous (over-count) census count for group a in area g.

We do not observe these quantities directly but if we did the level of over-count would be given by

$$\gamma_{ag} = \frac{X_g^{(a)}}{Y_g^{(a)}} = \frac{Y_g^{(a)} + E_g^{(a)}}{Y_g^{(a)}}. \qquad (1)$$

Therefore to estimate over-count we replace the quantities with estimates to give

$$\hat{\gamma}_{ag} = \frac{\hat{Y}_g^{(a)} + \hat{E}_g^{(a)}}{\hat{Y}_g^{(a)}}. \qquad (2)$$

Brown *et al* (2009) outlines the framework for this estimation in detail. The key is that any CCS return sampled in postcode p of area g that matches to a Census return in postcode p of area g contributes to Y in area g, while any CCS return sampled in postcode q of area g or any other area h that matches to a Census return in postcode p of area g contributes to E in area g. So $\hat{Y}_g^{(a)}$ is the weighted sum of the correct census returns identified by the CCS in area g while $\hat{E}_g^{(a)}$ is the weighted sum across the whole country of incorrect returns in area g identified by the CCS sampling the individual elsewhere.

To put this into operation we need to define a and g. In the basic under-count estimation, the grouping a is defined by a set of five year age-sex groups and g is either a single Local Authority (LA) or a small group of contiguous LAs, which form an Estimation Area based on achieving a certain sample size in the CCS . However, it is unlikely that the CCS will support over-count estimation at the same level. In addition, it is clear from previous research that students need to be treated as a separate category for over-count analysis. Over-count is also generally more prevalent amongst young adults, although there is less distinction between males and females. Therefore, our default 'a' can be broader age groups (combined across males and females) but with a separation for students amongst young adults[4]. Due to students the forming of the area grouping g needs thought. This is because students are highly clustered where they should be counted (in LAs with

[1] Type 2 only.
[2] Type 2 and Type 3.
[3] Area g will be an aggregation of LAs.
[4] We use a possible grouping within the simulations, see Table 2.

Universities) and when over-counted by parents they will tend to be dispersed (and often in an LA without a University). If we mix these LAs together then (2) will end-up down-weighting all students in the University LA to compensate for the student over-count in the non-University LA. Hard-to-count also helps as within an LA with a University the students attending the University will tend to be concentrated within the hard-to-count areas, while student home addresses will more likely be spread in the easier to count areas. Therefore, our default g will be two groups of LAs within each Government Office Region (GOR) based on whether they have a University or not, and then split by hard-to-count within each grouping. Splitting by 'University or not' is only necessary for the age group we split for students. In other age groups g can be just GOR by hard-to-count or we can utilise some alternative grouping of LAs.

4) Extending to Utilise the Match One Duplicate Estimation

We can define a new quantity $D_{hg}^{(a)}$, the number of erroneous (over-count) census returns for group a in area h within GOR g, which is a duplicate of the correct return within the same GOR. Match one is a focused sample of census returns to identify duplicates within a GOR but this cannot establish the correct return from the erroneous return. However it should measure with good accuracy $D_{g}^{(a)}$, the total number of erroneous duplicates with the correct return somewhere in the GOR, using the estimated number of matched pairs $\hat{P}_{g}^{(a)}$ found within the GOR. Using the CCS in the same way that we estimate the total number of erroneous returns E, outlined in section 3, we can also estimate $D_{hg}^{(a)}$ using the CCS elsewhere to highlight the erroneous half of the duplicate. To mirror match one this can be limited to CCS sample within the same GOR. Therefore, if the CCS samples postcode p within area j of GOR g and identifies the erroneous half of a duplicate within postcode q from area h of the same GOR g we have $D_{qhg,pjg}^{(a)} = 1$. Our estimate of $D_{hg}^{(a)}$ is then given by

$$\hat{D}_{hg}^{(a)} = \sum_{q \in h} \sum_{j \in g} \sum_{p \in CCS_j} w_{pjg} D_{qhg,pjg}^{(a)} \qquad (3)$$

where w_{pjg} is the CCS sampling weight for postcode p selected in area j of GOR g. We now sum over h to give $\hat{D}_{g}^{(a)} = \sum_{h \in g} \hat{D}_{hg}^{(a)}$, which should also equal $\hat{P}_{g}^{(a)}$. They will not be equal, but assuming match one gives a higher precision (and is less sensitive to any reporting bias that might exist within the CCS) we can improve the CCS based estimate by calibrating to give

$$\tilde{D}_{hg}^{(a)} = \sum_{q \in h} \sum_{j \in g} \sum_{p \in CCS_j} \tilde{w}_{pjg} D_{qhg,pjg}^{(a)} \qquad (4)$$

where the new weight is given by $w_{pjg} \times \hat{P}_{g}^{(a)} \Big/ \hat{D}_{g}^{(a)}$. This revised weight can then be used in the estimation of E to be used in (2), assuming the level of Type 2 over-count is correlated with all erroneous over-count (Type 2 and Type 3).

5) Simulating Over-count

To create over-count we take a single estimation area KO (Coventry and Solihull) and treat this as a closed population. Therefore, a CCS within the area can identify both Y and E.

The basic estimation simulations described by Brown and Sexton (2009) provide a response status for everyone in the estimation area (this is then sampled to give an appropriate CCS during the simulation process). The standard version of the simulation system provides a response flag of 1 to 4. A second response flag is produced to indicate if an over-count occurred for that individual. The crucial point here is that we do not attempt to place the over-count record elsewhere so it appears as over-count in the same place. Therefore, if the CCS samples the real record it will identify the over-count record to contribute to estimating E regardless of whether the real record is in the Census and contributing to Y. This simplifies the simulation compared to the reality of searching outside the CCS sample to find E. The flags are shown in Table 1.

Table 1: Standard and over-count flags

Standard flag	Modified (over-count) flag	Contribution to Census only count	Contribution to CCS only count	Contribution to 'both' (in Census and CCS) count	Over-count
1	1	0	0	1	No
2	2	1	0	0	No
3	3	0	1	0	No
4	4	0	0	0	No
1	5	1	0	1	Yes
2	6	2	0	0	Yes
3	7	1	1	0	Yes
4	8	1	0	0	Yes

Over-count is introduced into the KO population in targeted population groups. These are chosen based on previous research, and at levels in order to give approximately 0.5% over-count, overall, on the number of individuals in KO. These groups and introduced levels of over-count are given in Table 2.

Table 2: Over-count groups

Group	Characteristics	Level of over-count
1	Anyone not contained in groups 2 to 5 (see below)	0.1%
2	Persons aged 3 to 17	0.25%
3	Students aged 18 to 24	12%
4	Non-Student aged 18 to 24	0.5%
5	Persons aged 85+	0.25%

400 simulations have been run on KO. As this is a simulation, we can construct the level of over-count using (1) based on the whole area with the CCS being treated like a second census as well as estimate the level of over-count using (2) when the CCS is only a sample. Figure 1 shows the variation of the estimated propensities for over-count over the 400 simulations in each case for the five groups identified in Table 2. From Figure 1 we can see that both ways of estimating the level of over-count have distributions centred on the values given in Table 2 for each of the groups. However, as we move to the sample based estimate (CCS) there is, as we would expect, a considerable increase in variability. Also, the higher the rate of over-count, the larger the variability in our estimate of the level of over-count; something that we also see with increasing levels of under-count.

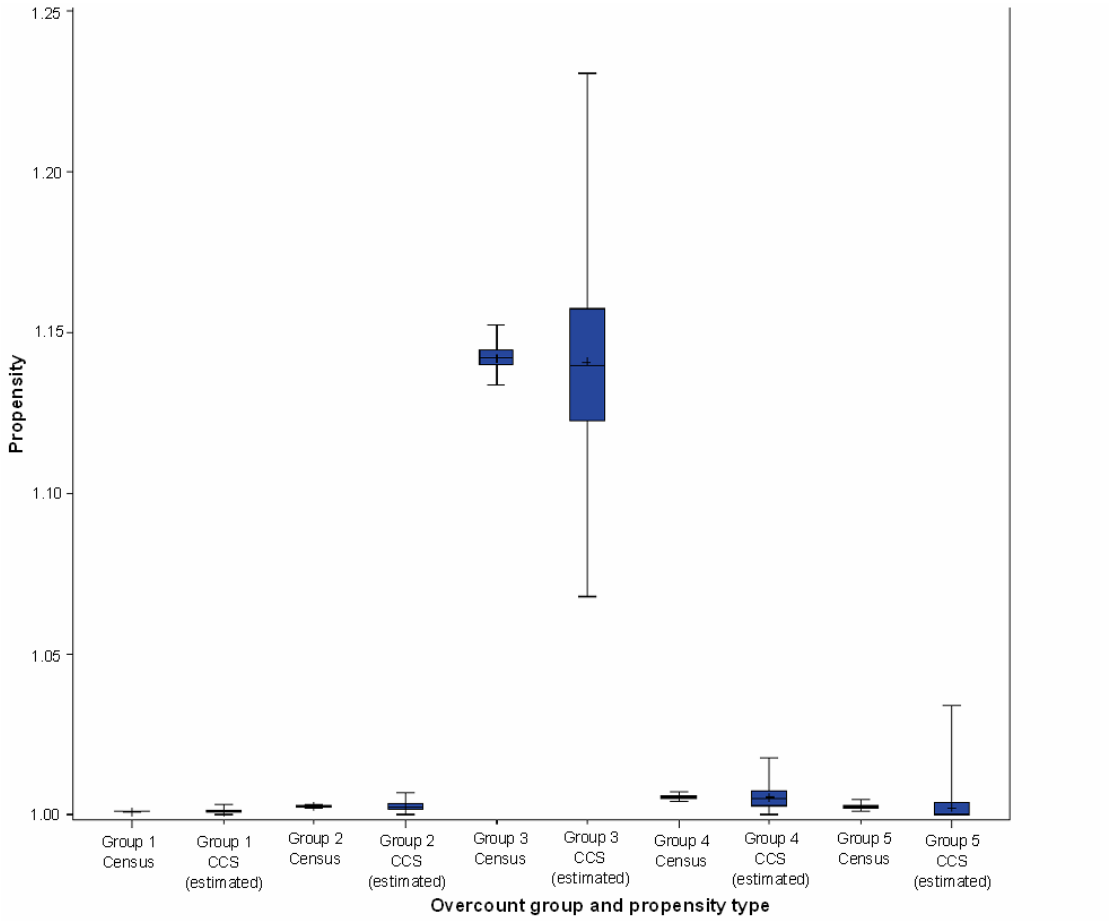

Figure 1: Box-plot of propensity calculated from the simulated Census and CCS, by over-count group.

6) Adjusting the DSE for the Estimated level of Over-count

The basis of the under-count estimation is the use of dual-system estimation at the level of a cluster of sample postcodes p within groups a. We defined the dual-system estimator (DSE) as $\hat{N}_{ap} = \frac{Z_{ap} \times Y_{ap}}{M_{ap}}$ where Z_{ap} is the CCS count for group a in cluster p, Y_{ap} is the census count of correct individuals, and M_{ap} is the matched count. However, we observe X_{ap}, the census count including over-count, so we would adjust the DSE to give $\tilde{N}_{ap} = \frac{Z_{ap} \times X_{ap} / \hat{\gamma}_{ag}}{M_{ap}}$ as suggested by Brown and Abbott (2008). This recognises that we cannot adjust the census count directly but given the underlying over-count propensity is constant across area g then $E\left[X_{ap} / \hat{\gamma}_{ag} \mid X_{ap} \right] \cong Y_{ap}$ and therefore $\tilde{N}_{ap}$ is approximately unbiased for $\hat{N}_{ap}$ given the observed counts Z, X, and M.

The actual estimation strategy goes further and applies the Chapman correction to the DSE as we are applying the DSE to small populations. This gives the basic estimator as

$$\hat{N}^{C}_{ap} = \frac{(Z_{ap}+1)\times(Y_{ap}+1)}{(M_{ap}+1)} - 1 \qquad (5)$$

and as a first order approximation it can be written as

$$\hat{N}^{C}_{ap} \cong \hat{N}_{ap}\left(1+\frac{1}{Z_{ap}}+\frac{1}{Y_{ap}}\right)\left(1-\frac{1}{M_{ap}}\right)-\left(1-\frac{1}{M_{ap}}\right). \tag{6}$$

As with the DSE, we will observe X_{ap} rather than Y_{ap} so the adjusted DSE with Chapman correction, which we implement, is given by

$$\tilde{N}^{C}_{ap} = \frac{\left(Z_{ap}+1\right)\times\left(X_{ap}\big/\hat{\gamma}_{ag}+1\right)}{\left(M_{ap}+1\right)}-1 \cong \tilde{N}_{ap}\left(1+\frac{1}{Z_{ap}}+\frac{\hat{\gamma}_{ag}}{X_{ap}}\right)\left(1-\frac{1}{M_{ap}}\right)-\left(1-\frac{1}{M_{ap}}\right) \tag{7}$$

and therefore $\tilde{N}^{C}_{ap}$ is approximately unbiased for $\hat{N}^{C}_{ap}$ given the observed counts Z, X, and M. In fact we extend (7) further to recognise that the groupings, a, used in the estimation of under-count (typically five year age-sex groups) do not match those used to estimate (2), the level of over-count, so our final adjusted DSE with Chapman correction has the form

$$\tilde{N}^{C}_{ap} = \frac{\left(Z_{ap}+1\right)\times\left(X_{a_1p}\big/\hat{\gamma}_{a_1}+X_{a_2p}\big/\hat{\gamma}_{a_2}+1\right)}{M_{ap}+1}-1 \tag{8}$$

where a_i is an over-count grouping that overlaps with under-count grouping a. Using the same results we can see the approximate unbiasedness will hold provided we have

$$E\left[X_{a_ip}\big/\hat{\gamma}_{a_ig} \mid X_{a_ip}\right] \cong Y_{a_ip} \tag{9}$$

for any group a_i and that over-count is operating independently of under-count.

7) Simulation Results for the Adjusted Estimator

The estimation system was run for all 400 simulations and the population estimates by age-sex group produced for 3 scenarios. These are as follows:

1. The estimated population without over-count – this is estimated using the standard 1 to 4 response status flags (as indicated in Table 1 of section 4) using the cluster level DSE with Chapman correction (3), and a simple ratio estimator;
2. The estimated population with over-count – this is estimated using the 1 to 8 modified response status with over-count flags (as indicated in Table 1 of section 4) using the cluster level DSE with Chapman correction (3), and a simple ratio estimator;
3. The estimated population with over-count and estimated propensity adjustment – this is estimated using the 1 to 8 modified response status with over-count flags (as indicated in Table 1 of section 4) using the cluster level DSE with Chapman correction adjusted based on the estimated level of over-count for each of the 5 over-count groups (6), and a simple ratio estimator.

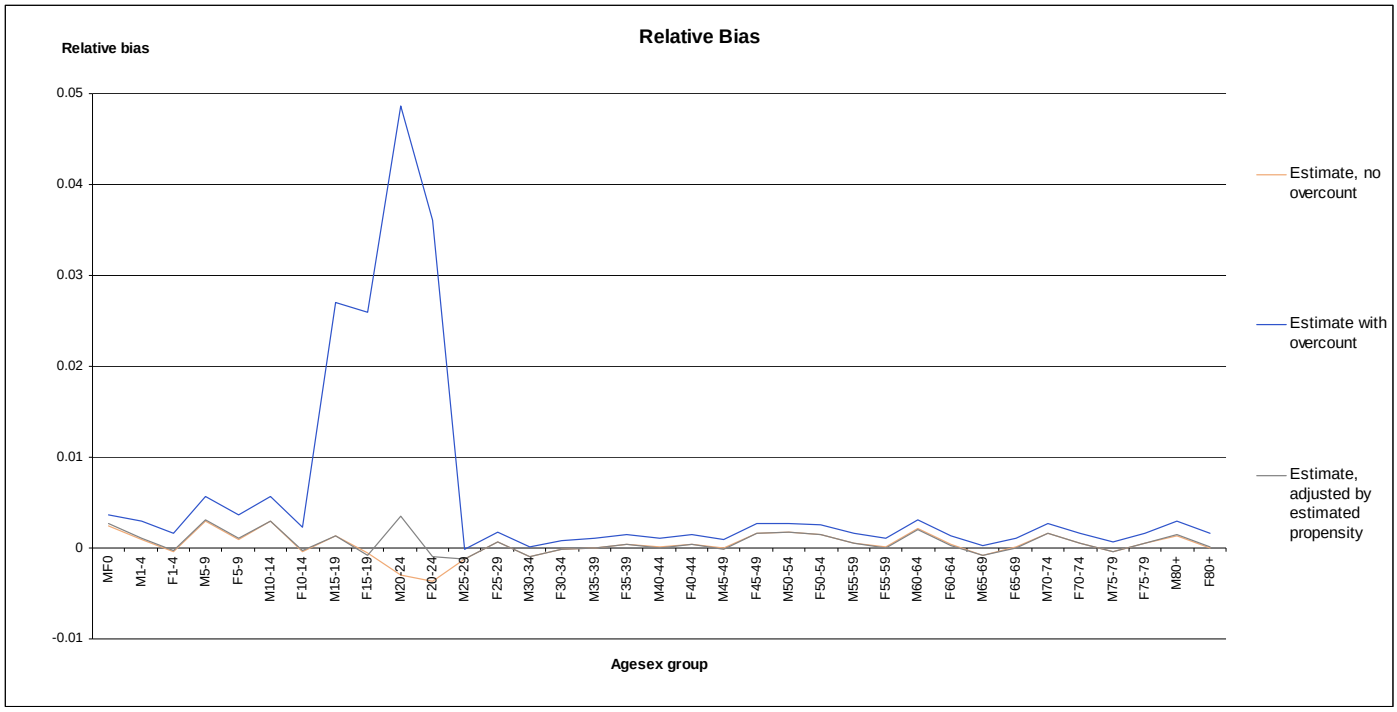

Figure 2: Relative bias across age-sex groups for the three scenarios.

Figure 2 shows the relative bias for the three scenarios. As we would expect, when we introduce the over-count (scenario two) but do not correct for it this induces a bias into the estimates, which is high in those groups that overlap with the over-count group for students. However, when we correct for the estimated level of over-count using the estimates shown in Figure 1 (scenario three) we move back to the performance of the estimator without over-count (scenario one). To see the comparison between scenario two and scenario three more clearly, Figure 3 just presents the bias for these two. This more clearly shows the groups where there is most deviation is in the male and female age 20 to 24 groups. This is likely to be due to substantial amount of over-count added into these groups by the student over-count group, which results in a much greater variability around the propensity estimate, as shown in Figure 1. For all other age-sex groups, the estimate produced using an estimated propensity adjustment is very close to the estimate produced without any over-count. This indicates that, generally, the over-count methodology proposed does work well in terms of getting back to our original population estimate.

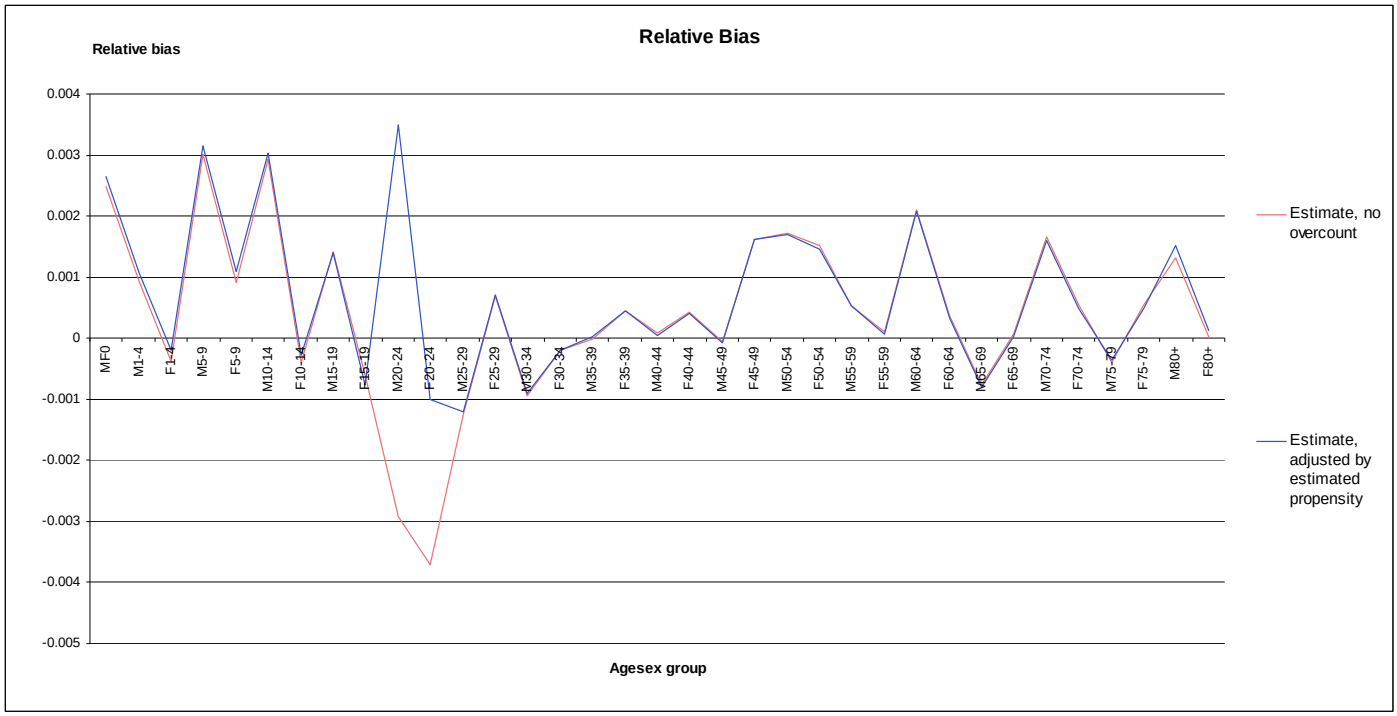

Figure 3: Relative bias across age-sex groups for scenario one (no over-count) and scenario three (over-count with an adjustment).

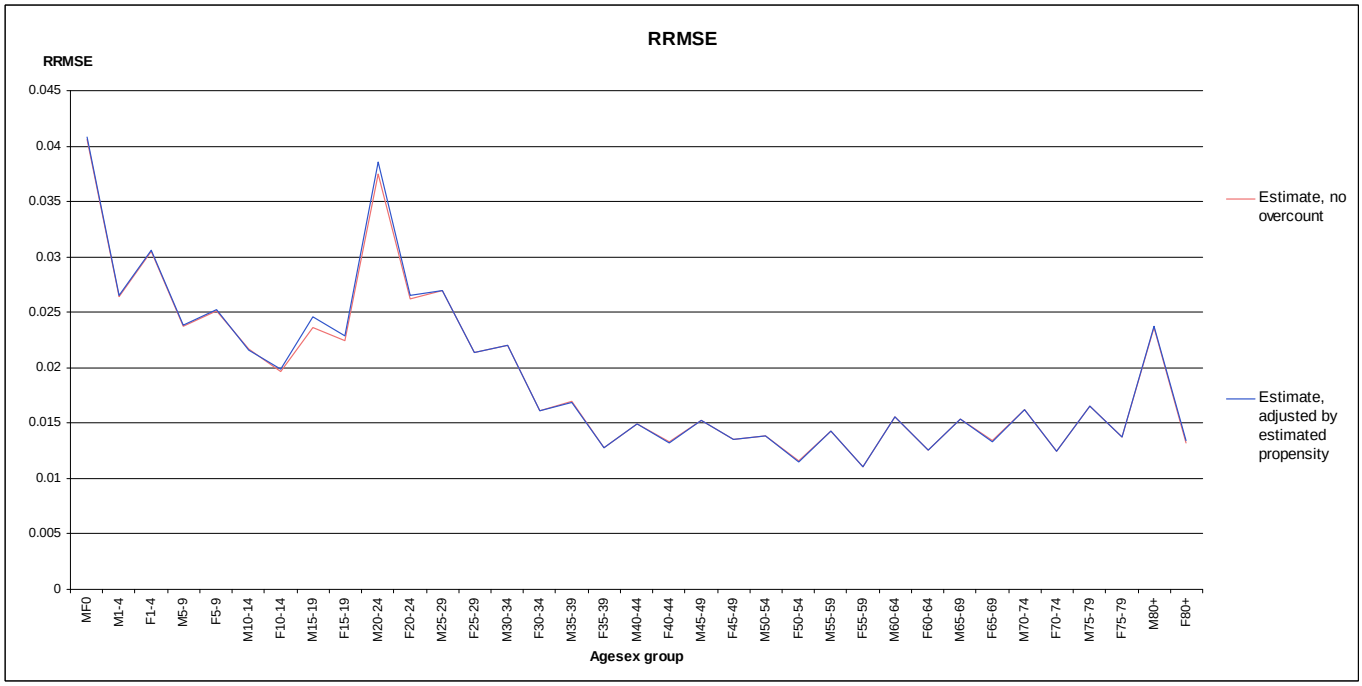

Figure 4: RMSE across age-sex groups for scenario one (no over-count) and scenario three (over-count with an adjustment).

We have seen in Figure 1 that the estimated propensity varies considerably relative to its level and so we need to consider the impact of having to adjust for over-count on the overall error given by the relative root mean square error (RRMSE). Figure 4 compares the RRMSE for the two scenarios shown in Figure 3. Figure 4 indicates that carrying out the estimation process using an estimated propensity adjustment introduces very minimal additional variance into the estimates. For those age-sex groups where there was a slightly higher propensity for over-count, and where the 5 over-

count groups did not align exactly with the age-sex groups used in the population estimates[5], the estimates adjusted for the estimated level of over-count are a little more variable, but this amount is very small.

8) Conclusion

It appears from these results that, generally, estimating the propensity from the CCS recreates the propensity in the overall population reasonably well, especially for groups with relatively low levels of over-count. Looking at the relative bias results for the adjusted population estimates, the over-count adjustment does manage to successfully remove the majority of the over-count, and is certainly a substantial improvement over not doing any kind of over-count adjustment at all. In addition, the results for the RRMSE indicate that the adjustment does not introduce a significant amount of extra variation into the final population estimates.

Based on these results we recommend that the approach to estimating and adjusting for over-count be implemented in the 2011 estimation.

9) Future Work

It may be worthwhile to produce a second set of results with increased levels of over-count. Currently the simulations contain approximately 0.5 per cent overall over-count. If over-count is a much bigger problem (as seen in the US), then levels closer to 4 per cent should also be investigated.

In order to more accurately demonstrate the complete methodology, a second set of simulations would also be recommended, where we use a second CCS sample within each simulation to estimate the level of over-count. This more closely reflects the fact that E is estimated from the CCS samples outside the area identifying the over-count within the area.

We have also proposed in section 4 (as well as Abbott and Brown for the NSMAC(13) (2008) and Brown and Taylor for the GSSMAC (17) (2009)) linking the external estimation of duplicates by matching a sample of Census returns to the Census (match one) with the CCS based estimate of over-count given in (2). This involves calibrating the CCS based estimate of duplicates to the external estimate through (4), which we expect to be less variable. However, given the small increase in the RRMSE seen in Figure 4, we may need further consideration of the benefit of this additional step.

10) References

Abbott, O. and Brown, J. (2007) Overcoverage in the 2011 UK Census. Paper presented to 13th Meeting of the National Statistics Methodology Advisory Committee. Available at www.statistics.gov.uk/methods_quality/downloads/NSMAC13_Census_Overcoverage.pdf

Abbott, O. and Large A. (2009) Measuring the level of duplicates in the 2011 Census. Paper presented to 17th Meeting of the GSS Methodology Advisory Committee. Available at www.statistics.gov.uk/methods_quality/downloads/GSSMAC17.pdf

[5] These groups are the males and females, age 15 to 19, and males and females aged 20 to 24. These four groups contribute to the over-count groups consisting of persons aged 3 to 17, students aged 18 to 24, and non-students aged 18 to 24.

Brown, J. and Abbott, O. (2008) Response to Discussion at NSMAC on Estimation of Over-Count. Note for 14th Meeting of the National Statistics Methodology Advisory Committee. Available at www.statistics.gov.uk/methods_quality/downloads/Estimation_overcount_followup.doc

Brown, J. and Sexton, C. (2009) Estimates from the Census and the Census Coverage Survey, Paper presented at the 14th GSS Methodology Conference. Available on request.

Brown, J., Taylor A. and Abbott O. (2009) Overcount in the 2011 Census: Estimation Issues. Paper presented to 17th Meeting of the GSS Methodology Advisory Committee. Available at www.statistics.gov.uk/methods_quality/downloads/GSSMAC17.pdf