

## ONE NUMBER CENSUS STEERING COMMITTEE

### Capture-recapture estimation in a One Number Census

1. Sections 1, 2 and 3 of this paper provide a basic introduction to capture-recapture methodology
2. Sections 4 and 5 look at the application of capture-recapture estimation to county populations using data from the Census, a Census coverage survey, and an administrative list. Problems with the basic Dual/Triple System Estimator are identified and explored in simulations.
3. Further work is planned to:
  - a) investigate using simulation the use of log-linear modelling techniques to look at interactions between sources;
  - b) investigate the proportion of the census non responses which have to be counted in the CCS for a reasonable response.
  - c) liaise with other researchers to investigate advanced matching techniques.
4. **Members of 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.**

**Wayne Codd  
Census Division  
Office for National Statistics**

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

**November 1997**

# Capture-recapture estimation in a One Number Census

James Brown, Wayne Codd, Lisa Buckner

## 1. Introduction

1.1 In order to estimate the population by age and sex at county level, it is intended that counts from the 2001 Census will be adjusted for underenumeration using a Census Coverage Survey (CCS) and estimates from administrative sources. The simplest way of doing this would be to consider the number of people enumerated only in the actual Census, and compare this with the total number of people found on any source. An undercoverage ratio could then be calculated and used to ‘uplift’ estimates for small areas, by age and sex.

1.2 This very basic estimator makes one crucial assumption; that everyone the Census should be counting appears on at least one of the sources. Unfortunately, this is unlikely to be the case, so there is an unknown population of people that are not found on any of the lists. The size of this population needs to be estimated. It is proposed that this is done using a form of capture-recapture modelling.

## 2. Lincoln-Peterson model

2.1 The principles of capture-recapture estimation can be best explained by considering the method in its simplest form, known as the Lincoln-Petersen model or Dual-System Estimation (DSE).

2.2 In terms of human population estimation, two lists which attempt to record the population of interest should be used. For example, the Census roll and the Census Coverage Survey (CCS) listing could be used.

2.3 The two lists then need to be matched in order to determine the number of people appearing on both lists, and on one or the other. These can be set out in a 2x2 contingency table, as in Table 1, with example counts.

	CCS record	No CCS record	
Census record	Both 765	Census only 135	Total Census records 900
No Census record	CCS only 85	Missed by both X	
	Total CCS records 850		Total population ?

**Table 1** Contingency table of matches and non-matches

2.4 The working assumption is that the percentage of Census records for which there is a corresponding CCS record is the same as the percentage of the whole population for which there is a CCS record. This is set out in Equation 1 below.

**Equation 1**

$$\frac{\mathbf{A} \quad \text{No. of Census records with matching CCS record}}{\mathbf{B} \quad \text{No. of Census records}} = \frac{\mathbf{C} \quad \text{No. of CCS records}}{\mathbf{D} \quad \text{County population}}$$

2.5 The Census can be pictured as a sample of the county population, from which the number of individuals captured by the CCS is to be estimated from a known population frame. However, the position here is the other way round - therefore the number of CCS records is known, not the sampling frame (county population). Therefore, instead of estimating C in Equation 1, as usual in statistical inference, it is D that is to be estimated.

2.6 Equation 1 can be re-written to derive various counts.

$$\text{County population} = \frac{\text{No. of Census records} \times \text{No. of CCS records}}{\text{No. of matching records}}$$

$$\text{Number missed by both registers} = \frac{\text{No. on CCS only} \times \text{No. on Census only}}{\text{No. of matching records}}$$

2.7 Using the example counts from Table 1, the estimates would be calculated as follows:

$$\text{County population} = \frac{900 \times 850}{765} = 1000$$

$$\text{Number missed by both registers} = \frac{135 \times 85}{765} = 15$$

**3. Assumptions**

3.1 The Lincoln-Petersen model relies on certain crucial assumptions holding true:

**i) Independence**

The probability of an individual appearing on the CCS listing should not be related to their appearance on the Census.

If the probabilities of an individual being recorded by each of the lists are related, as seems likely, A would again be higher, and the unknown population underestimated.

## ii) *Homogeneity*

All individuals are equally likely to appear on a given list. This is a particular property of the individual and may depend on age, sex, geography or social status.

It seems certain that heterogeneity of capture probabilities exists between individuals, violating this assumption. In equation 1, this would make A higher, and the left-hand-side of the equation higher, and therefore D lower. The size of the unknown total population will be underestimated as we will be misled into thinking that a higher proportion of people are being recorded on both lists.

## iii) *Closure*

Members of the population do not die, and there are no births, between the Census and the CCS.

The data from the two registers should be collected on dates as close to each other as possible, to uphold the assumption of closure. This assumption can be made more reliably for human populations than for animal populations whose birth and death rates are higher. In addition, field-workers for the CCS can at least check the status of households with respect to their location on Census date.

## iv) *Accurate matching*

It is important that individuals who appear on both listings are correctly identified and successfully matched. If they are not, A will be too small and the unknown population will be over-estimated.

## v) *No list-inflation*

The registers should be 'clean', with no individual being registered twice, or deceased individuals remaining on the register.

If the registers are not totally clean, C would be artificially high and the unknown population would be over-estimated.

## 4. Two list simulation

4.1 The aim of the first set of simulations is to assess how the simple dual-system estimator performs when the fundamental assumption of independence between the two lists is violated. To introduce dependence between the Census and CCS counts the odds ratio (R) was used. This is defined as:

$$R = \frac{P_{11}P_{00}}{P_{10}P_{01}}$$

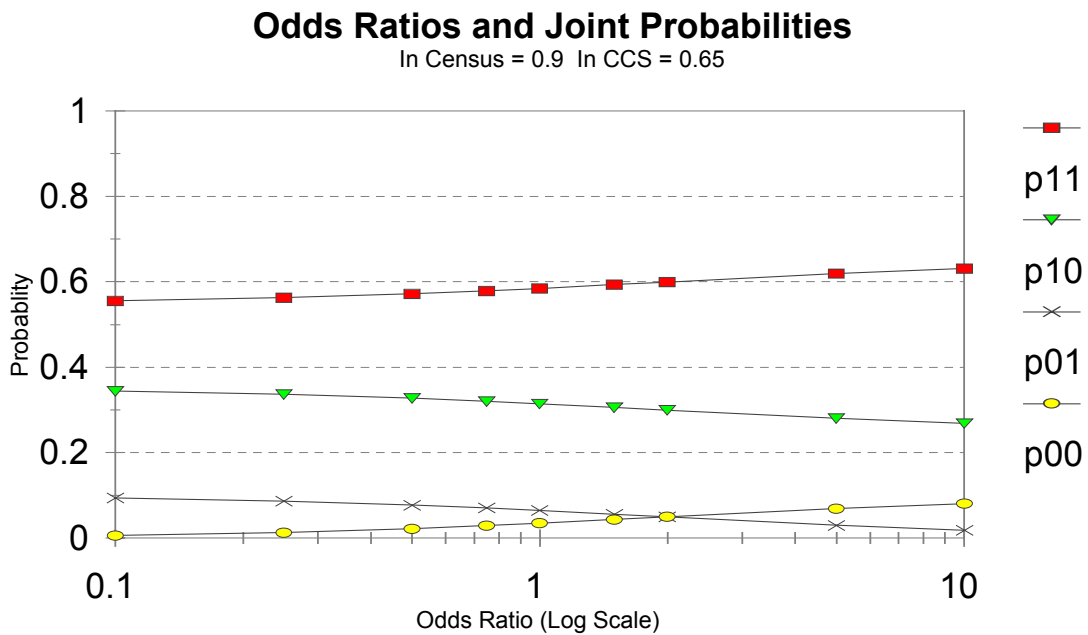
where  $p_{11}$  = probability of being counted in both listings

$p_{10}$  = probability of being counted by the Census but not the CCS

$$\begin{aligned}
&= p_{1+} - p_{11} \\
p_{01} &= \text{probability of being counted by the CCS but not the Census} \\
&= p_{+1} - p_{11} \\
p_{00} &= \text{probability of being missed in both listings} \\
&= p_{0+} - p_{01} = 1 - p_{1+} - p_{+1} + p_{11} \\
p_{1+} &= \text{probability of being counted by the Census} \\
p_{+1} &= \text{probability of being counted by the CCS}
\end{aligned}$$

4.2 An odds ratio equal to one signifies independence between the sources. It varies on a log scale. Ratios less than one imply that people missed by the Census have greater ‘odds’ of being in the CCS than the people who were counted in the Census. In other words  $p_{11}$  goes down. Ratios greater than one imply that people counted by the Census have greater ‘odds’ of being in the CCS than the people who were missed in the Census. This effect is shown in Figure 1. In other words  $p_{11}$  goes up. In the first case the DSE overestimates and in the second it underestimates.

**Figure 1**



4.3 Given the odds ratio for the 2x2 table and the marginal probabilities of being counted in the Census and being counted in the CCS it is possible to calculate all the cell probabilities (see Annex A).

4.4 A population of 1000 people was simulated and each person was given a probability of being in the Census and another of being in the CCS. The odds ratio was fixed and each person was assigned their four cell probabilities. For each person the CCS probability is the same but for the Census it varies around some fixed value,

say 0.9. To generate the two lists a multinomial trial was carried out for each individual. The possible outcomes are:

- i) 11 - in both Census and CCS;
- ii) 10 - in the Census but not in the CCS;
- iii)01 - not in the Census but in the CCS and
- iv)00 - not in the Census or the CCS.

4.5 From the two lists the DSE was used to estimate the total population (truth=1000). This was repeated 1000 times for each simulation. A series of simulations was carried out to investigate the result of varying the dependence using the odds ratio, as well as the marginal coverage of each list.

### ***Two list simulation results***

4.6 The results from the simulations are summarised in Figures 2 and 3 below, demonstrating the effects of dependence when Census Coverage equals 90% and 70% respectively. A CCS coverage of 85% is expected generally, but may be lower for some types of people. We have therefore simulated a range of CCS coverages.

4.7 Figure 2 shows that the estimator behaves as expected when dependence is varied using the odds ratio. In addition, the estimator is sensitive to the coverage of the CCS. As coverage falls, the effect of the dependence increases. At the extremes of dependence, for odds ratios of 0.2 and 5, the estimator performs very poorly. For these odds ratios, a 95% confidence interval (estimated over the 1000 simulations) does not include the 'truth'.

4.8 Figure 3 shows a similar pattern as Figure 2, however, the effects of dependence and poor CCS coverage are exaggerated by the lower Census coverage of 70%.

4.9 These simulations show that when coverage of both lists is high, the effect dependence can exert on the accuracy of the estimator is limited. However, as coverage falls, and the independence assumption of the Lincoln-Petersen model is violated, the Dual-System estimator fails.

Figure 2

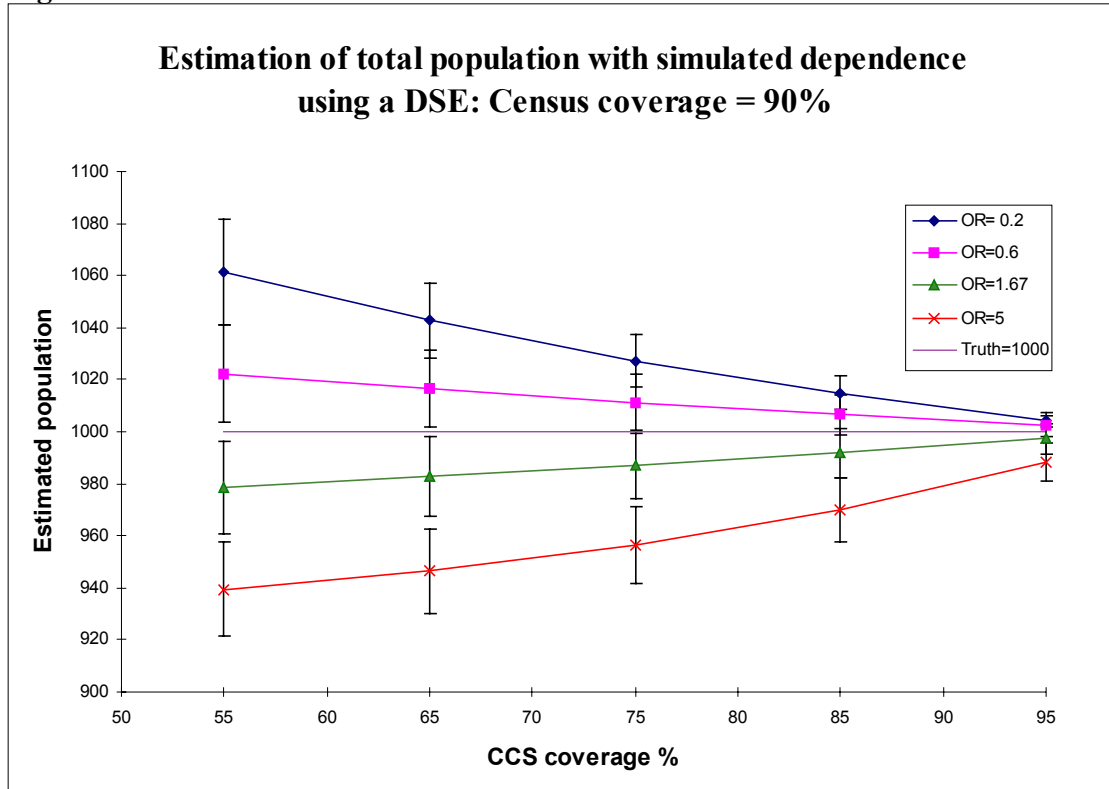
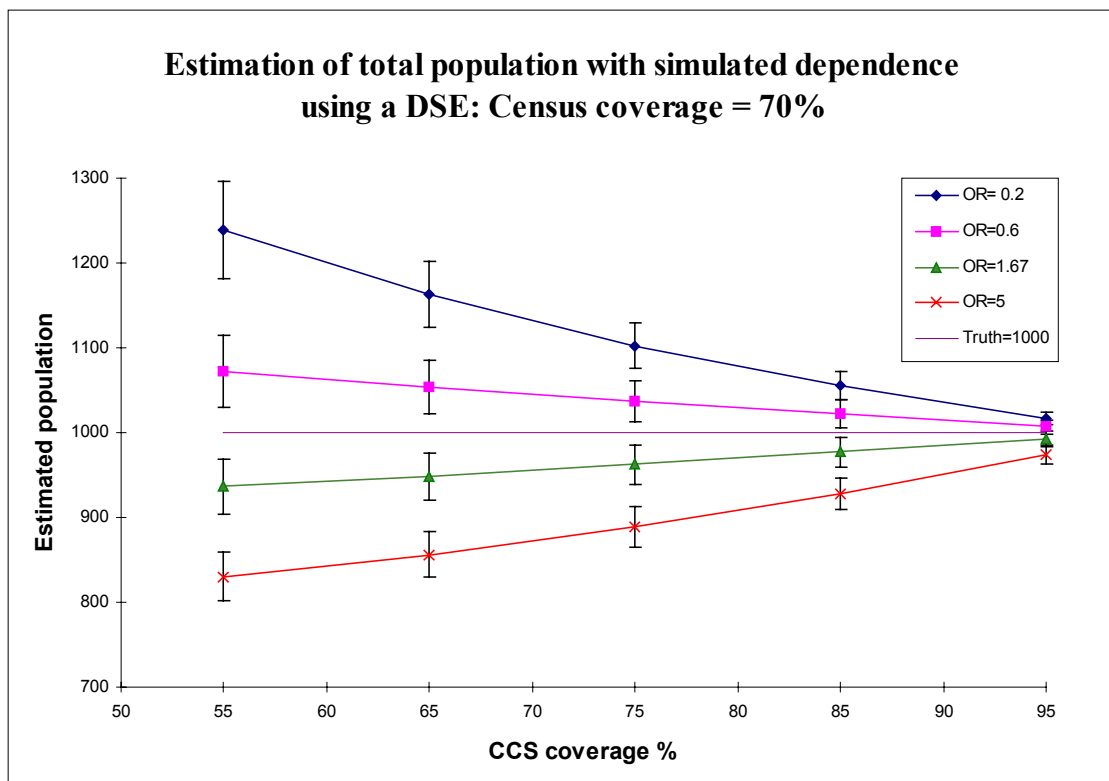


Figure 3



## 5. Triple System estimation (TSE)

5.1 The principles of the Lincoln-Petersen model can be extended to use three sources. This may be desirable to improve the precision of the estimate when the assumption of independence between the Census and the CCS is violated. Using a National Health Service register such as the FHSA listings (now known as HAs), we can construct Table 2 and estimate the cell  $X_{000}$  (those missed from all three lists). It should be noted, however, that when using administrative data of this type as the third source, it is important to be particularly aware of the assumptions of ‘no list-inflation’ and ‘accurate matching’.

		FHSA		No FHSA		
		CCS	No CCS	CCS	No CCS	
Census	Y	$X_{111}$	$X_{101}$	$X_{110}$	$X_{100}$	$X_{1++}$
	N	$X_{011}$	$X_{001}$	$X_{010}$	$X_{000}$	$X_{0++}$
		$X_{+11}$	$X_{+01}$	$X_{+10}$	$X_{+00}$	$X_{+++}$

**Table 2** 3-way contingency table

From Table 2:

$$X_{+1+} = X_{+11} + X_{+10} = \text{Total number in the CCS}$$

$$X_{++1} = X_{+11} + X_{+01} = \text{Total number on the administrative list}$$

5.2 Using Table 2, and assuming independence between all three lists, the Triple System Estimator can be calculated as follows (details of this are presented in Annex B).

$$\text{Total} = \sqrt{\frac{\text{Census} \times \text{CCS} \times \text{FHSA}}{\text{Number in all}}}$$

### *Simulation using three lists*

5.3 An independent administrative list was added to the two-list simulation. Instead of the DSE, the TSE was used which still assumed independence between all three lists. To generate the three lists, the two lists model was extended such that the Census and CCS outcomes were still dependent but the joint outcome from the Census and CCS was independent of the administrative list. This situation is considered to be quite realistic when the administrative list used is a National Health Service registration list or something similar. The public are less likely to see registering with a General Practitioner in the same light as data collected by the government, which may be regarded as existing for the purpose of recording their characteristics and/or movements.

5.4 Again, for a given odds ratio between the Census and CCS, along with the three marginal probabilities, it is possible to calculate all eight cell probabilities. These probabilities were computed for each of the 1000 people in the simulated population. The above simulation was then repeated, but this time an eight way multinomial trial was used to determine the state of an individual on all three lists, i.e. 111, 101, 110, or 100 being 'In Census' outcomes etc.

5.5 From the three lists the total population was estimated and for each run this was repeated 1000 times. The whole simulation was repeated for various list coverages and odds ratios.

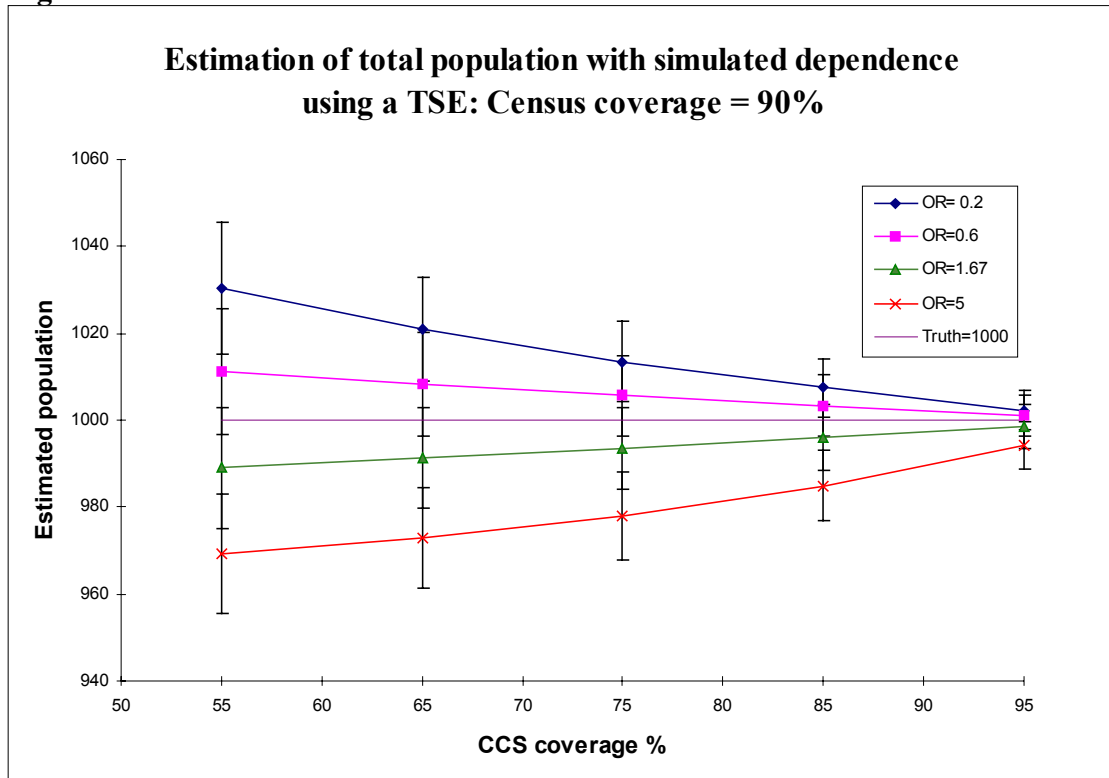
5.6 The final stage introduced list-inflation on the administrative list. This represents people who have been registered twice, died or moved out of the area but have not been removed from the list. This was achieved very simply, by inflating the value of cell 001 - people on the list only, such that the total population given by the list was 102% of the true population (of which 90% were in the true population).

### ***Three list simulation results***

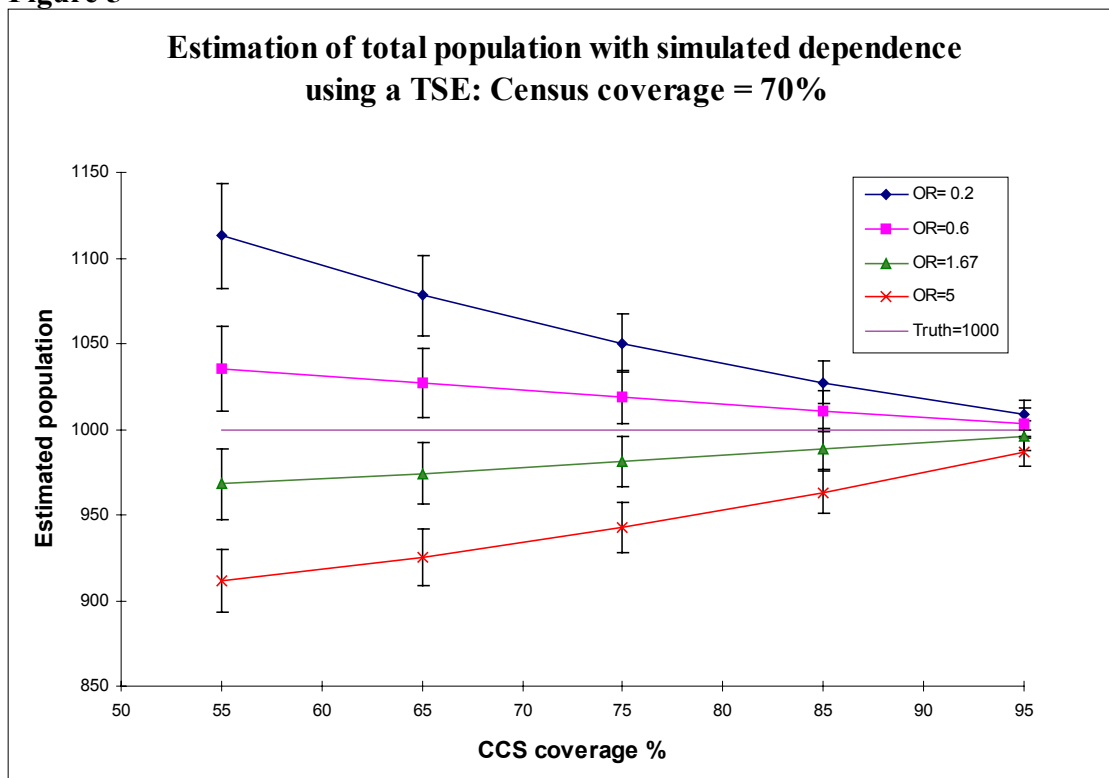
5.7 The results from the simulations are summarised in Figures 4 to 7 below, with and without list-inflation, and again with Census coverage set at 90% and 70% respectively.

5.8 Figures 4 and 5 show a similar pattern to the DSE in Figures 2 and 3. Again, the effects of dependence and poor CCS coverage are exaggerated by the lower Census coverage of 70%. However, in the case of the TSE simulations the estimate of the total population is greatly improved by the use of a third list which is independent of the Census and CCS. The introduction of this list reduces the error in the estimates by approximately 50%. When coverage of the Census or the CCS is high, the estimates are as accurate as 980 or 1010 for high degrees of dependence.

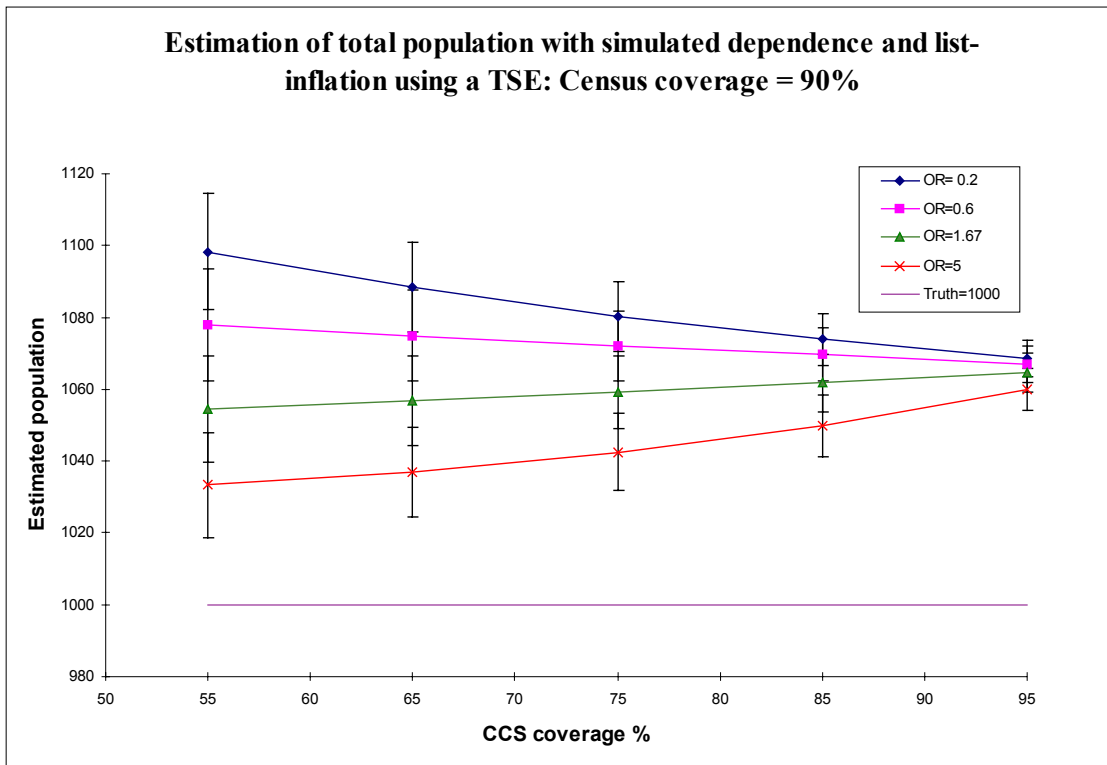
**Figure 4**



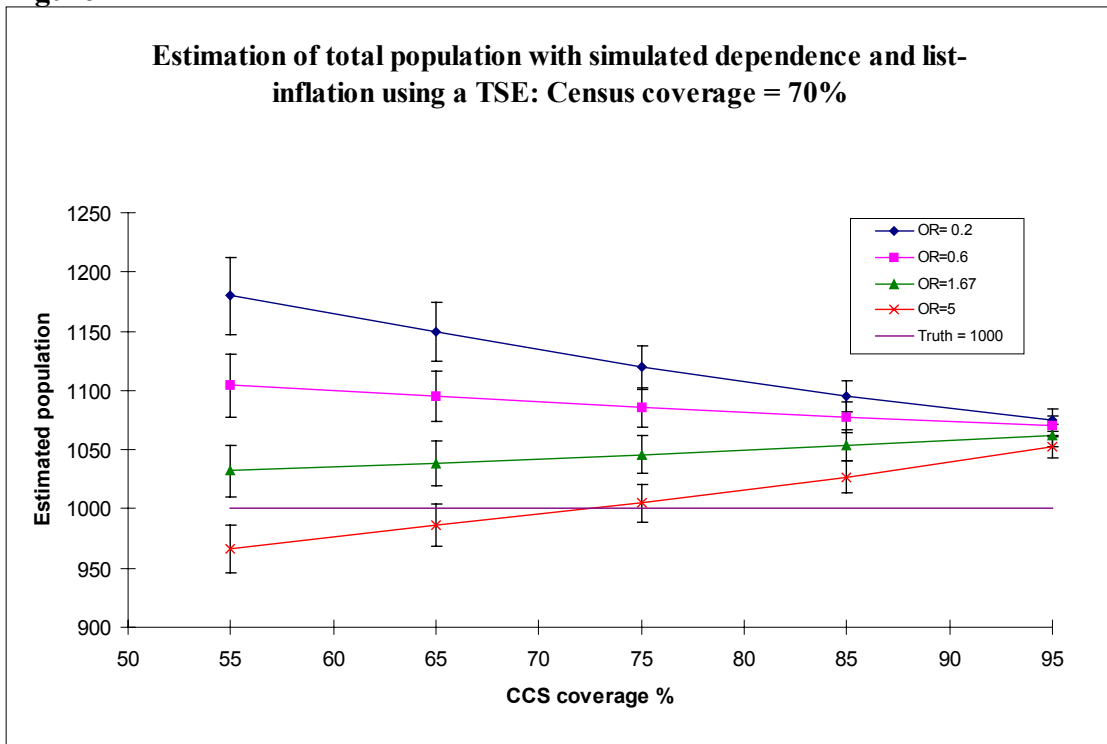
**Figure 5**



**Figure 6**



**Figure 7**



5.9 Figures 6 and 7 illustrate the effect of the presence of list inflation on the third independent list. When compared with Figures 4 and 5 it can be seen that list inflation has the effect of inducing positive bias on the estimation of the total population. It effectively leads to an upward ‘shift’ in the estimate therefore Figures 6

and 7 are similar in pattern to Figures 4 and 5 but lines are 'shifted' vertically upwards.

5.10 The list-inflation of 2% introduced into the simulations results in an increase in the over-estimation of the total population of approximately 200% for the case of the 90% Census coverage, 55% CCS and for odds ratio = 0.2 (1030 increases to 1098). The effect is reduced for the case when the Census coverage is 70% (1112 increases to 1179). For the situation when negative bias exists, that is when the odds ratio is equal to 5 say, the effect of the positive bias due to the list inflation on the third list results in a low net error in the estimate of the total population. However, the gross error in this case is great and this cannot be overlooked.

## **6. Conclusions**

6.1 The simulations have shown that the Dual-system estimator performs better when coverage of both sources is high, and reduces the effect of dependence. At low levels of coverage, dependence between the Census and the CCS biases the estimate severely.

6.2 The addition of an independent third list improves the accuracy of the estimate but finding a suitable source is difficult since the problems of list-inflation outweigh any benefits derived from the use of a third list. The magnitude of the actual list-inflation present in the listings needs to be investigated.

## **7. Areas for further research**

### ***Log-linear models***

6.3 Research has been undertaken in the US to overcome the problems of applying capture-recapture estimation to human populations. By using a third list, various models have been developed in two papers by Darroch *et al.* (1993) and Zaslavsky and Wolfgang (1993) which use log-linear modelling of cell counts to model two-way interaction (and therefore dependence) between sources. The choice of model depends on the structure of data in the contingency table. These complex log-linear models are not possible with just two lists.

6.4 The basic results presented in this paper will be compared with those obtained from the loglinear models.

### ***Matching accuracy***

6.5 The problem of matching has not been addressed in this paper, but is fundamental to the success of any capture-recapture methodology. If the real matches between any sources cannot be identified in a sizeable number of cases then the problems of dependence and list-inflation are insignificant in comparison. The use of names in the matching process is crucial to linkage between census and Health Authority sources, particularly if HA records are not up to date and individuals need to be found registered in other Authorities. As explained in ONS(ONC(SC))97/14, the improved success given by linkage using name will be investigated empirically. In addition, the development of high-quality matching techniques will be researched.

## *Dependence*

6.6 Further work will be undertaken to compare Dual-system and Triple-system estimators with different levels of dependence between the lists used.

6.7 Empirical work will be carried out to determine the expected dependence between the lists.

## **8. References**

- Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993)**  
*A three-sample multiple-recapture approach to Census population estimation with heterogeneous catchability.* Journal of the American Statistical Association, **88**:1137-1148.
- Zaslavsky, A. M. and Wolfgang, G. S. (1993)**  
*Triple-system modelling of Census, post-enumeration survey, and administrative-list data.* Journal of Business & Economic Statistics, **11**:279-288.

## ANNEX A - Calculation of the cell probabilities

The following two-way table shows the individual cell and marginal probabilities for being counted and not counted by the Census and CCS.

	In CCS	Not in CCS	
In Census	$p_{11}$	$p_{10}$	$p_{1+}$
Not in Census	$p_{01}$	$p_{00}$	$p_{0+}$
	$p_{+1}$	$p_{+0}$	1

Here:

- $p_{11}$  = probability of being counted in both listings
- $p_{10}$  = probability of being counted by the Census but not the CCS  
=  $p_{1+} - p_{11}$
- $p_{01}$  = probability of being counted by the CCS but not the Census  
=  $p_{+1} - p_{11}$
- $p_{00}$  = probability of being missed in both listings  
=  $p_{0+} - p_{01} = 1 - p_{1+} - p_{+1} + p_{11}$
- $p_{1+}$  = probability of being counted by the Census
- $p_{+1}$  = probability of being counted by the CCS

From this table the odds ratio R is defined as:

$$\text{Odds ratio } R = \frac{p_{11}p_{00}}{p_{10}p_{01}}$$

For fixed odds ratio R the aim is to obtain the individual cell probabilities. Given the marginal probabilities the other cell probabilities can be expressed in terms of  $p_{11}$  and these marginal probabilities, thus:

$$\begin{aligned} p_{10} &= p_{1+} - p_{11} \\ p_{01} &= p_{+1} - p_{11} \\ p_{00} &= p_{0+} - p_{01} = 1 - p_{1+} - p_{+1} + p_{11} \end{aligned} \quad (\text{A1})$$

Substituting these into the odds ratio gives:

$$R = \frac{p_{11} \times (1 - p_{1+} - p_{+1} + p_{11})}{(p_{1+} - p_{11})(p_{+1} - p_{11})}$$

Rearranging this gives:

$$p_{11}^2(1 - R) - p_{11}((p_{1+} + p_{+1})(1 - R) - 1) - R \times p_{1+} \times p_{+1} = 0$$

This can be solved as a quadratic in  $p_{11}$  using the formula:

$$p_{11} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where ,

$$-b = (p_{1+} + p_{+1})(1 - R) - 1$$

$$a = (1 - R)$$

$$c = -R \times p_{1+} \times p_{+1}$$

The choice of +ve or -ve root is constrained by  $0 \leq p_{11} \leq 1$

This gives  $p_{11}$  in terms of fixed marginal probabilities and a fixed odds ratio. Using the formula in (A1) it is possible to calculate the other three cell probabilities.

## ANNEX B - Derivation of the Triple System Estimator.

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

$$P(C = 1) = \text{Probability of being counted in the census} = X_{1++}/X_{+++}$$

$$P(S = 1) = \text{Probability of being counted in the CCS} = X_{+1+}/X_{+++}$$

$$P(A = 1) = \text{Probability of being on the administrative list} = X_{++1}/X_{+++}$$

Under independence:

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

Using the notation of Table 2, the above can be written as:

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

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

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

i.e.

$$\text{Total} = \sqrt{\frac{\text{Census} \times \text{CCS} \times \text{FHSA}}{\text{Number in all}}}$$