



ONS(ONC(SC))98/14

## **ONE NUMBER CENSUS STEERING COMMITTEE**

### **One Number Census Matching**

1. This paper describes the background to the 2001 Census / CCS record matching initiative and provides a brief description of record matching using probability weights. Examples are provided of the calculation of these probability weights and the effect of different cut-off weights on the matching results. Areas of future work are outlined and the results that can be expected discussed.
2. **Members of the Steering Committee are asked to:**
  - a) **note the paper;**
  - b) **comment and advise on the approach outlined, either at the meeting on 13 November 1998 or in writing by 27 November 1998.**

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

## ONC MATCHING EXECUTIVE SUMMARY

This paper describes the work carried out to match the Census and CCS data.

### **S1. Background**

S1.1 Underenumeration is a fact of censal life. Estimation of underenumeration is problematic due to its differential nature. In the 1991 Census, although the overall level of coverage was high (estimated at nearly 98%), some population subgroups suffered from an estimated undercount of as much as 23% (young males in some cities).

S1.2 The 1991 Post Enumeration Survey was designed to assess both the quality of the census data and the coverage of the census. In 2001 the issue of underenumeration will be addressed explicitly. Approximately three weeks after census night the Census Coverage Survey (CCS) will be undertaken in an attempt to identify all individuals living within the sampled postcode units. The CCS sample has been designed such that accurate estimates of underenumeration should be possible for all census areas.

S1.3 The ability of the CCS to identify wholly missed households and people missed from counted households is obviously crucial to the success of the ONC, but equally important is to have a census/CCS matching process that correctly identifies these extra households and people. Based on the experience in 1991 we are looking to find the missing 2% and while the absolute numbers may be large, percentages are small. Thus the ONC process requires a matching methodology that does not introduce as many errors as it is aiming to resolve.

### **S2. Overview**

S2.1 The different and independent enumeration methodologies employed by the Census and CCS mean that simple matching using a unique identifier common to both lists is not possible. Furthermore, simple exact matching on the variables collected in common by both methods is out of the question as there will be errors in both sets of data caused by incorrect recording, misunderstandings, the time gap between the Census and CCS, errors introduced during processing etc. The size of the CCS also means that hand matching is not feasible. Thus a largely automated process involving probability matching is necessary.

### **S3. Probability Matching**

S3.1 Probability matching involves assigning a probability weight to a pair of records based on the level of agreement between the two records. The probability weights reflect the likelihood that the two records correspond to the same individual. Two records may then be assigned as a match, even if they disagree on a small number of details, provided the probability weight exceeds a pre-determined threshold.

S3.2 Firstly, blocking variables must be identified. A blocking variable, e.g. postcode, is used to reduce the number of comparisons required by an initial grouping of the records. Probability matching is only undertaken within blocks as defined by the blocking variable.

S3.3 Matching variables such as name, tenure and month of birth are then compared for each pair of records within a block. Provided the variables being compared are independent of each other, the probability weights associated with each variable can be summed to give an overall probability weight for the two records. Records are matched if the likelihood of them relating to the same individual exceeds an agreed threshold.

#### **S4. Deriving the Weights**

S4.1 Before the probability weights can be calculated it is necessary to perform a clerical matching exercise to produce a file of matched records. This file should be representative of the total set of records to be matched. The file is then used to attain the likelihood that a pair of outcomes is observed when the records belong to the same person, for each possible pair of outcomes.

#### **S5. Strategy**

S5.1 The CCS collects data for two purposes; to enable the data to be matched against the census; and to identify the characteristics of underenumeration via the modelling process, so that adjustments can be applied to the whole population. In order that the second part is not biased by the first the matching and modelling variables should be as independent as possible.

S5.2 As the data are structured both geographically and by individuals within households we utilise this structure within the matching strategy.

S5.3 The key stages of the matching are as follows:

- 1) Produce a clerically matched file
- 2) Derive the matching weights
- 3) Use 'blocking' variables to reduce the number of comparisons made
- 4) Match households
- 5) Match individuals within matched households
- 6) Clerically check any CCS forms left unmatched

#### **S6. Simulation Data**

S6.1 A simulated census and census coverage survey were used to test the matching strategy described above. Systematic errors were introduced into the simulated data sets to allow for the demonstration of the calculation of the probability weights. Although the principles and practice of probability matching can be demonstrated using these simulated data sets, it is not possible to assess the accuracy of the ultimate matching process as this is dependent on the quality of the actual 2001 Census and CCS data. Further research will address this.

#### **S7. Results**

S7.1 The results of the pseudo matching exercise were assessed for various cut-off weights. These are the probability weights below which no positive match will be assigned. The simulation showed that the false match rate decreases as the cut-off threshold increases. This demonstrates that probability weights can be used to distinguish, to some degree, between true and false matches.

S7.2 It should be noted that many of the matches missed in 2001 would be resolved by clerical matching. However, it is desirable to reduce the size of this clerical matching exercise as far as possible as it will be both costly and time consuming.

S7.3 The programmes used to run the automated matching took approximately one hour to match the 22,000 households and the individuals contained within them. This shows that the automated matching procedure can be completed quickly. It is the post-automated matching clerical match that will prove time consuming and hence the extent of this must be minimised.

## **S8. Matching Errors and the Dual System Estimator**

S8.1 It is necessary to consider the effect that errors in the matching will have on the dual system estimates of the population. It is shown in this paper that a false match rate of  $\alpha\%$  will similarly give rise to an error in the dual system estimate of approximately  $\alpha\%$ . This is the error that is directly attributable to inaccuracies in the matching and not any other source.

## **S9. Further Work**

- Simulated census and CCS to be used to assess the impact of data quality on the matching procedure.
- Methodology for estimating false match rates to be developed.
- Decide on an appropriate programming language. Investigate the accuracy and efficiency of the matching programmes.
- Investigate the contribution of individual's names and address names/numbers to the accuracy of the matching procedure.

## **S10. Conclusions**

**This paper:**

- **shows that it is possible to use probability weights to distinguish between true and false matches**
- **demonstrates the implementation of the proposed matching strategy. Although a number of issues remain to be resolved, the basic principles of the matching have been established**
- **demonstrates that, assuming a reasonable level of accuracy in the census and CCS data, the matching of the data sets is achievable within an acceptable time frame**

## ONE NUMBER CENSUS MATCHING

### 1. Introduction

1.1 Census data are used for the allocation of Central and Local Government funds for the following ten years. It is therefore of great importance that these figures are accurate down to small areas and subgroups of the population.

1.2 Despite all efforts to maximise coverage it is very likely that there will be a degree of underenumeration. Estimation of underenumeration is problematic due to its differential nature. In the 1991 Census, although the overall level of coverage was high (estimated at nearly 98%), some population subgroups suffered from an estimated undercount of around 20% (young males in some cities).

1.3 The 1991 Post Enumeration Survey was designed to assess both the quality of the census data and the coverage of the census. In 2001 the issue of underenumeration will be addressed explicitly through the One Number Census (ONC) process. The major tool will be a post enumeration survey, known as the Census Coverage Survey (CCS). This will be a large postcode based survey carried out soon after the census and will involve the re-enumeration of a sample of postcodes. The CCS sample has been designed such that accurate estimates of underenumeration should be possible for all census areas. This process is described fully in the steering committee paper ONS(ONC(SC))98/16.

1.4 A key requirement of the ONC process is the need to match accurately the data collected in the CCS with those collected in the census so as to identify those households and individuals who do not appear on a census form. Even a few matching errors may be of critical importance, since population adjustments can be less than 1% in some instances (Jaro (1989)).

### 2. Introduction to Automated Record Linkage

2.1 The objective of this paper is to describe the record linkage strategy that will be used in the ONC process.

2.2 Record linkage, or probability matching, involves assigning a probability weight to a pair of records based on the level of agreement between the two records. The probability weights reflect the likelihood that the two records correspond to the same individual. Two records may then be assigned as a match, even if they disagree on a small number of details, provided the probability weight exceeds a pre-determined threshold.

2.3 When designing a statistical record matching procedure it is important to bear in mind the reasons for which the files are being linked. If files are linked using proper mathematical models, then the files can be analysed using statistical models such as regression and loglinear models (Scheuren & Winkler (1993)).

2.4 It is also necessary to consider some of the basic principles of probability matching; simplicity and flexibility. Beyond a certain level, increasing complexity of the weight calculation routine tends to involve diminishing returns. Record linkage is not about the mechanical application of complex and abstract rules. It is about the sensitive and flexible application of very simple basic rules (Kendrick (1997)).

2.5 Simplicity can be achieved by careful consideration of the variables to be used in the matching process and also by the choice of sensible blocking variables. Blocking variables are used to restrict the numbers of comparisons made. Records are only compared where they have identical blocking variables. Since errors may occur in blocking variables it is usually necessary to have a number of blocking variables, independent as far as possible, each of which will be used in turn as selection criteria for proposed pairs of records.

2.6 Blocking variables should be chosen so that they bring together the majority of true pairs, whilst comparing very few false pairs. The selection of blocking variables is an important aspect of the matching process. The literature on probability matching has tended to focus primarily on the calculation of probability weights. However, it is suggested by Kendrick *et al.* (1998), improvements in the accuracy of a matching process are just as likely to be achieved by sensible consideration of the blocking variables used.

2.7 Simplicity can also be achieved by minimising the number of variables used in the matching routine. Independence between these matching variables removes the need for complex interaction calculations. In principle, any items whose level of agreement or disagreement influences the probability that two records do or do not belong to the same person can be used by the computer algorithm. However, items should be statistically independent as far as possible (Copas & Hilton (1990)).

### **3. Matching Strategy**

3.1 It is proposed that there will be two separate matching exercises. The first will match households and the second individuals within matched household pairs.

#### ***Household level matching***

3.2 Any matching strategy will clearly be dependent on the variables common to both sets of data. Once these have been established it is necessary to decide which variables to use as blocking variables, which as matching variables and which to disregard for the matching process. Consideration should be given to the level of independence between blocking, matching and analysis variables.

#### ***Household Variables***

3.3 The possible matching variables are:

- Postcode
- Address name / number
- Tenure
- Individual's names
- Individual's sex
- Individual's dates of birth
- Individual's marital status
- Individual's ethnicity

3.4 Of these, address name / number and individual's names may not be available on the final 2001 Census data sets. A decision on the capture of these variables will be made after the Census Dress Rehearsal.

3.5 Proposed additional derived variables are:

- Household structure
- Head of Household

3.6 An algorithm for the derivation of Head of Household is given in Appendix 1. Any head of household definition must be highly likely to select the same individual from the census data as from the CCS data. For this reason it is not possible simply to select the first listed individual. This algorithm requires detailed analysis during the 1999 Census Dress Rehearsal to ensure that it fulfils its required properties.

3.7 The algorithm for deriving household structure has not yet been developed. It is not possible to derive this variable at present, as the form of the relationship question has not yet been finalised. However, this paper has been written on the assumption that a suitable definition can be applied to the census and CCS data.

### ***Household Blocking Variables***

3.8 The first requirement is to decide on suitable household blocking variables. At least two sets of blocking variables are required so that records that are not matched on the first pass of the data are compared on different criteria for the second pass. Ideally these blocking variables should be independent as far as possible to ensure that the accuracy of the first set of blocking variables does not affect the accuracy of the second. The variables selected should also minimise the number of pairwise comparisons made.

3.9 The sampling units for the CCS are postcode units. Therefore it is sensible to use full postcode as a blocking variable. However, postcode on its own will not sufficiently reduce the number of pairwise comparisons being made. Therefore, for the first pass of the data, it is proposed to block also on address name / number.

3.10 It is possible for a postcode to be recorded incorrectly and it may also be difficult, in some circumstances, to determine where one postcode ends and another begins. Therefore, for the second pass of the data, it is suggested to use postcode and all contiguous postcodes as one blocking variable. For the second part of the blocking variable something independent of address name / number is required. It is proposed to use an encryption (e.g. Soundex Code) of surname of head of household.

3.11 Any household records still unmatched after these two passes of the data would then be checked clerically to determine if an appropriate match can be found.

### ***Household Matching Variables***

3.12 Table 2 below summarises the possible household matching variables.

<b>Matching Variable</b>	<b>Comment</b>
Postcode Unit	Proposed blocking variable
Address name / number	Proposed blocking variable
Tenure	Tenure may prove unable to distinguish sufficiently between households. This should be tested in the Census Dress Rehearsal.
Individual's Name	Proposed blocking variable
Individual's Sex	Analysis Variable
Individual's Date of Birth	Age is an analysis variable. Age will also be dependent between household members. However, day and month of birth should provide independent matching variables if the information collected is sufficient.
Individual's Marital Status	This will be dependent on age and the marital status of other individuals in the household
Individual's Ethnicity	Analysis Variable.
Household Structure	This may prove to be a poor indicator of a match. It is proposed to test this during the Census Dress Rehearsal. Household structure is also dependent on the number of people within a household. This is being used as a modelling variable.
Head of Household	Will probably be derived from age, sex and activity last week variables.

**Table 2: A summary of possible household matching variables.**

3.13 Ideally, the matching procedure would be based on mutually independent matching variables, distinct from the proposed blocking and analysis variables, which have relatively small numbers of discrete possible outcomes. This would simplify the calculation of the matching weights and reduce the size of the clerically matched file required before automated matching can be performed.

3.14 The values taken by matching variables, such as tenure, may provide good distinguishing characteristics at a high level, but not within a postcode unit. It may therefore be advisable to have different matching weights for different sampling units. One suitable way of attaining this is to have different probability weights for each Hard to Count classification. Another possibility is to train the matching weights for each postcode using the Estimation and Maximisation (EM) algorithm. These methods will be tested during the Census Dress Rehearsal.

3.15 Regardless of which method of weight allocation is ultimately selected, it is clear that simplicity in the weight calculation is desirable. If the EM algorithm is chosen, simplicity becomes increasingly necessary due to time and computing limitations.

3.16 Consideration of the possible matching variables, as given in Table 2, leads to the following initial selection:

- Tenure
- Household Structure
- Surname of Head of Household

3.17 These variables may be assumed to be independent.

3.18 Further work may lead to the selection of items of individual level information as household matching variables. This issue is currently under consideration.

3.19 The matching and blocking variables are clearly dependent on an accurate method of determining the head of household. If this algorithm does not identify the same individual in the majority of census and CCS cases, the accuracy of the matching process will be affected. The head of household algorithm will be tested following the Census Dress Rehearsal.

### ***Individual level matching***

3.20 Individual level matching will be performed within matched household pairs. All possible pairings of individuals will be considered, with the most likely combination being selected as the matched individuals.

3.21 For some household pairs it may not be possible to allocate matches for any of the individuals contained within these households. In these cases it is proposed to re-classify these households as unmatched.

### ***Individual Matching Variables***

3.22 The possible individual level matching variables are:

- Forename
- Surname
- Sex
- Day of Birth
- Month of Birth
- Year of Birth
- Marital Status
- Ethnicity

3.23 Of these, first name and surname may not be available in the final 2001 Census datasets. The contribution to the accuracy of the matching process made by these variables will be assessed following the 1999 Census Dress Rehearsal.

3.24 In addition to the variables listed above it is proposed to derive a 'Relationship to Head of Household' variable. This algorithm cannot be fully specified at present, as the relationship question has not yet been finalised.

3.25 Blocking variables are not considered here since all comparisons are within matched household pairs. For large households it may be necessary to block on some matching variables, but this issue has not been addressed at present.

3.26 The need for independence between matching and analysis variables leads to the following being suggested as individual level matching variables:

- Forename
- Surname
- Day of Birth
- Month of Birth
- Relationship to Head of Household

3.27 The suitability of these variables, their independence and ability to distinguish between matches and non-matches, will be carefully investigated during the 1999 Census Dress Rehearsal.

3.28 The matching process described above has been converted into a set of matching algorithms, given in Appendix 2. These algorithms detail the steps taken by the computer programmes written to perform the matching exercise.

#### **4. A Simulation of the Matching Process**

4.1 Following the 1997 Census Test, a test of the Census Coverage Survey was performed. This exercise did not provide sufficient data to demonstrate the derivation of the probability weights and the full matching procedure described in this paper. Therefore it was decided to simulate a census and CCS large enough to illustrate the processes being proposed here.

4.2 The 1997 Census Test data were taken as the true population and records selected to imitate the taking of a census and CCS. Selection probabilities were used to reflect the undercoverage possible in a true Census and CCS. Errors were then introduced into the data sets so that the principles of probability matching on data containing errors could be demonstrated.

4.3 It is not possible to test the matching procedure in full using these data as they have several limitations. The relationship matrix from the 1997 Test was not coded onto the data set. Therefore there is no means of deriving a household structure or relationship to head of household variable. Also, the errors introduced into the data have all been systematic and are not necessarily realistic. The accuracy of the matching procedure is clearly dependent on the quality of the data and the errors present. It is important to note that, although these simulated data can be used to demonstrate the matching process, it is not possible to predict the accuracy of the 2001 automated match from these synthetic data. This assessment may be performed after the Census Dress Rehearsal, however this will still not give an entirely accurate estimate of the accuracy in the 2001 Census due to the compulsory nature of the census and increased national publicity nearing the census proper. These issues are discussed further in section 7.

4.4 The matching variables used for the simulated data sets are as follows:

4.5 For household level matching:

- Tenure
- Number of people in household
- Soundex code of surname of Head of Household

4.6 For individual level matching:

- First name
- Surname
- Day of Birth
- Month of Birth
- Marital Status

## **5. Calculation of Probability Weights**

5.1 In order for the matched data sets to be suitable for subsequent statistical analysis it is necessary for the matching to have been performed in a statistically rigorous manner, using suitable probability weights.

5.2 All of the appropriate methods of calculating probability matching weights are dependent on the existence of a large clerically matched file that can be assumed to contain a random sample of true matches. This file is then used to calculate the probability weights associated with each of the matching variables. Various methods of calculating probability weights are available. The two methods used here are:

- Likelihood ratios
- Hit-Miss Model

5.3 The likelihood ratio method of calculating probability weights is applicable to variables with a small number of discrete outcomes. The probabilities are based on the frequencies of each possible pair of outcomes in the clerically matched file and the marginal probabilities of each individual outcome. The methodology is described in more detail in Appendix 3. Examples of matching probabilities derived using likelihood ratios are given later.

5.4 The hit-miss model is applicable to discrete variables with a slightly larger number of possible outcomes. It assumes that, for each variable, there is a fixed probability of being recorded correctly. If that variable is recorded incorrectly, it takes each possible outcome with a probability proportional to the frequency of that outcome in the clerically matched file. The standard form of this model assumes that the probability of observing a blank record in both sets of data being matched is essentially zero. When this is not the case it is possible to extend the model and take the probability of a blank record as a random effect. This model may be further expanded to allow for differing frequencies of blank records in the two files being matched.

5.5 The methodology is described in detail in Appendix 4. Examples of probabilities derived using the hit-miss model are given below.

### ***Example of likelihood ratio probability weight calculation***

5.6 The simulated census used here contains 42,211 households and 102,181 individuals. The simulated CCS is based on a sample of the postcodes contained within the simulated census and consists of 22,977 households and 54,319 individuals. Each record on the simulated datasets contains a unique marker on which the files can be linked to give true matching pairs. This information was used to link the data to remove the necessity for a large clerical exercise to match a sample of the two sets of data. This produces a file of 22,781 matched household pairs from which to illustrate the calculation of probability weights.

5.7 Tenure has seven possible outcomes, including non-response. For illustrative purposes, the matching probability weights for tenure are here calculated using likelihood ratios. The frequency of each pair of outcomes in the simulated data set is shown in Table 3 below.

Frequency		CCS Tenure							Total
		Owns	Buying with Mortgage	Part rent part mortgage	Rents	Lives rent free	Other	Blank	
Census Tenure	Owns	6,378	57	-	-	-	-	41	6,476
	Buying with Mortgage	-	7,199	-	-	-	-	46	7,245
	Part rent part mortgage	-	-	155	-	-	-	3	158
	Rents	-	-	-	4,765	-	-	31	4,796
	Lives rent free	-	-	-	5	295	-	-	300
	Other	-	-	-	-	-	58	-	58
	Blank	198	340	2	378	13	1	2,816	3,748
	<b>Total</b>	<b>6,576</b>	<b>7,596</b>	<b>157</b>	<b>5,148</b>	<b>308</b>	<b>59</b>	<b>2,937</b>	<b>22,781</b>

Table 3: Table showing the frequency of tenure responses in the simulated census and CCS for true matched pairs of households.

5.8 It can be seen from Table 3 that many of the possible pairs of responses have no observed records. It is thought unlikely that so many of these cells would be empty in reality. Since it is not possible to calculate realistic log-likelihood matching weights when so many of the observed cells are zero, it was decided, for illustrative purposes, to adjust the above frequency table so that each row and column total remained the same, but each cell contained at least one observation. This adjusted frequency table is given as Table 4.

Frequency		CCS Tenure							Total
		Owns	Buying with Mortgage	Part rent part mortgage	Rents	Lives rent free	Other	Blank	
Census Tenure	Owns	6,374	57	1	1	1	1	41	6,476
	Buying with Mortgage	1	7,195	1	1	1	1	45	7,245
	Part rent part mortgage	1	1	150	1	1	1	3	158
	Rents	1	1	1	4,760	1	1	31	4,796
	Lives rent free	1	1	1	5	290	1	1	300
	Other	1	1	1	1	1	52	1	58
	Blank	197	340	2	379	13	2	2,815	3,748
	<b>Total</b>	<b>6,576</b>	<b>7,596</b>	<b>157</b>	<b>5,148</b>	<b>308</b>	<b>59</b>	<b>2,937</b>	<b>22,781</b>

Table 4: Table showing the frequency of tenure responses in the simulated census and CCS for true matched pairs of households, adjusted so that each cell contains at least one observation.

5.9 The above frequencies can be used to calculate the probability matching weights for each possible pair of observed outcomes. These probabilities are calculated as the natural log of each individual frequency, multiplied by the total number of cases, divided by the row and column totals for that frequency (see Appendix 3). For example, the probability weight for an observed pair consisting of 'Buying with a mortgage' in the CCS and 'Owns outright' in the census is given by:

$$\ln\left(\frac{57 \times 22,781}{6,476 \times 7,596}\right) = -3.635.$$

5.10 Table 5 shows the full set of probability matching weights. The magnitude of the probability weight indicates the strength of the evidence for or against the match. The sign of

the probability weight indicates the direction of the evidence, positive for a match and negative against. It can be seen from Table 5 that observing a pair of records where both have the tenure value ‘other’ gives the strongest evidence in favour of a match. Similarly, observing a pair of records claiming to ‘Own outright’ in the CCS and to be ‘Buying with a mortgage’ in the Census offers the strongest evidence against a match.

Log-Likelihoods		CCS Tenure						
		Owens	Buying with Mortgage	Part rent part mortgage	Rents	Lives rent free	Other	Blank
Census Tenure	Owens	1.227	-3.635	-3.798	-7.289	-4.472	-2.820	-3.014
	Buying with Mortgage	-7.646	1.091	-3.911	-7.401	-4.584	-2.932	-3.033
	Part rent part mortgage	-3.820	-3.964	4.925	-3.575	-0.759	0.894	-1.915
	Rents	-7.233	-7.377	-3.498	1.480	-4.172	-2.519	-2.993
	Lives rent free	-4.461	-4.605	-0.726	-2.607	4.270	0.252	-3.655
	Other	-2.818	-2.962	0.917	-2.573	0.243	5.847	-2.012
	Blank	-1.703	-1.302	-2.558	-0.804	-1.360	-1.580	1.762

Table 5: Table showing the probability weights for each possible pair of tenure responses.

5.11 Matching weights calculated using a log-likelihood approach are given in Appendix 5 for the number of people in a household and surname of head of household at the household matching level.

5.12 For individual level matching weights, log-likelihoods were used for marital status, fore and surnames. These weights are also given in Appendix 5.

**Example of hit-miss model probability weight calculation.**

5.13 Month of birth has thirteen possible outcomes, including non-response. The calculation of probability weights for month of birth is shown here to demonstrate the use of the hit-miss model.

5.14 The true link markers on the simulated data set were used to create a sample file of 10,547 true linked individuals. This file was used in place of the clerically linked file that would be necessary with operational data.

5.15 Of 10,547 pairs of records, 921 (9%) months of birth are blank on both files. The frequency of blank months of birth is 10.5% and 9.7% in the simulated census and CCS respectively. It was therefore decided to use the random effects version of the Hit-Miss model, allowing for differing frequencies of blank records in the two files being matched. Of the 9,338 pairs of records that have month of birth recorded, nine disagree. The resulting parameters for the hit-miss model are given in Table 6. For an explanation of these parameters see Appendix 4.

Parameter	$b_1$	$b_2$	$\hat{\sigma}_b$	$\hat{c}$								
Estimate	0.105	0.097	0.077	0.001								
Parameter	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$
Estimate	0.090	0.080	0.090	0.087	0.089	0.086	0.083	0.076	0.078	0.080	0.077	0.084

Table 6: Parameter estimates for the hit-miss model for month of birth in a sample of true matched individuals from the simulated census and CCS.

5.16 It is now straightforward (see Appendix 4) to estimate the probability of observing any pair of outcomes using the fitted model. Due to the sparseness of the data, the 13x13 table of observed and expected frequencies has been summarised in Table 7. In this table  $i$  is the month of birth in the CCS data and  $j$  is the month of birth in the census.

	Frequencies for the following values of x:												
	1	2	3	4	5	6	7	8	9	10	11	12	Blank
<i>Agreements</i>													
$i=j=x$	841	752	844	811	831	805	771	703	720	747	721	783	921
Expected	837.2	749.1	841.1	810.1	830.3	806.7	771.2	705.8	724.0	748.1	719.1	785.0	920.8
<i>Disagreements</i>													
$i=j+x$	0	0	0	1	1	0	2	0	0	1	0		
$i=j-x$	0	1	2	0	0	0	0	1	0	0	0		
Expected	0.8	0.8	0.7	0.6	0.5	0.5	0.4	0.3	0.2	0.2	0.1		
<i>Blank</i>													
$i=\text{blank}, j=x$	7	8	8	6	10	8	7	8	11	10	5	14	
Expected	9.2	8.2	9.2	8.9	9.1	8.8	8.4	7.7	7.9	8.2	7.9	8.6	
$i=x, j=\text{blank}$	11	11	12	19	14	21	19	19	17	15	14	14	
Expected	16.7	15.0	16.8	16.2	16.6	16.1	15.4	14.1	14.5	14.9	14.4	15.7	

**Table 7: Table comparing the estimated frequencies from the hit-miss model with those observed. Here  $i$  is the month of birth observed in the CCS and  $j$  is the month of birth observed in the Census.**

5.17 Combining the disagreement observed and expected frequencies to create two cells,  $j > i$  and  $j < i$ , provides a table where the expected value in each cell is greater than 5 and to which a  $\chi^2$  test can be applied. The value of  $\chi^2$  is 18.3 on 24 degrees of freedom (39 cells, 15 parameters), indicating an acceptable fit.

5.18 It is now possible to calculate log-likelihood ratios given the fitted model. This gives a log-likelihood of -6.644 for all differing pairs where neither record is blank, -1.681 for all differing pairs where the census record is blank, -2.193 for all differing pairs where the CCS record is blank and 2.149 where both records are blank. For pairs of records that are non-blank and agree the log-likelihoods are shown in Table 8 below.

Month of Birth	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Log-likelihood	0.428	0.539	0.424	0.461	0.437	0.465	0.510	0.599	0.573	0.541	0.580	0.493

**Table 8: Log-likelihoods for record pairs showing the same month of birth, based on the hit-miss model.**

5.19 The hit-miss model was also applied to day of birth. The resulting probabilities are given in Appendix 6.

### ***Results of Matching the Simulated Census and CCS***

5.20 The matching algorithms were performed on the simulated census and CCS with a range of different cut-off weights, below which weight a positive match will not be assigned. The accuracy of the matching procedure could then be assessed by comparing the household and individual identifiers to see which of the assigned matched pairs were true pairs and which false.

5.21 It was decided to use household cut-off weights such that all possible pairs of records would be considered at the household level. Matched pairs would then be rejected if no linked pairs of individuals were selected within the household. Table 9 below shows the results of the matching procedure for various individual level cut-off weights.

Level	Cut-off Weight	Total Possible matches	Matches Achieved	False Positive Matches	Percentage of Matches Missed	False Match Rate	Number of Matches Missed
<i>Household</i>	-	22,120	22,137	677	3.0%	3.1%	660
<i>Individual</i>	0	51,168	48,821	1,112	6.8%	2.3%	3,459
<i>Household</i>	-	22,120	21,934	554	3.3%	2.5%	740
<i>Individual</i>	4	51,168	48,184	864	7.5%	1.8%	3,848
<i>Household</i>	-	22,120	21,870	544	3.6%	2.5%	794
<i>Individual</i>	8.04	51,168	47,920	811	7.8%	1.7%	4,059
<i>Household</i>	-	22,120	19,673	3	11.1%	0.0%	2,450
<i>Individual</i>	8.05	51,168	44,376	13	13.3%	0.0%	6,805

**Table 9:** Table showing the results of the automated matching procedure on the simulated census and CCS for various cut-off weights.

5.22 Table 9 clearly shows a decreasing false match rate as the cut-off threshold increases. This demonstrates that the weight algorithm can be used to distinguish, to some degree, between true and false matches. The false match rates and missed matches are in no way representative of the figures expected in the actual 2001 Census / CCS matching process as these will be entirely dependent on the data quality in the two data sets.

5.23 One feature of the simulated data set is that it contains a large number of records where all individual level matching variables are blank. The probability weight attached to a pair of individual's records where the households have been matched but all individual level matching variables are blank is 8.047. The differences in the percentage of matches missed and the false match rate when these blank pairs of records are assigned as matches or non-matches can be seen from the Table 8.

5.24 It should be noted that many of the matches missed in 2001 would be resolved by clerical matching. However, it is desirable to reduce the size of this clerical matching exercise as far as possible as it will be both costly and time consuming.

5.25 The programmes used to run the automated matching took approximately one hour to match the 22,000 households and the individuals contained within them. This shows that the automated matching procedure can be completed quickly. It is the post-automated matching clerical match that will prove time consuming.

## 6. Matching Errors and the Dual System Estimator

6.1 It is necessary to consider the effect that false matches will have on the dual system estimates (DSEs) of the population. Two types of false matches are possible:

- False positive matches occur when two records relating to different individuals are assigned as a positive match.
- False negative matches occur when two records relating to the same individual are classified as a non-match.

This section considers the absolute numbers of matches achieved, irrespective of the relative numbers of false positive and negative matches. This absolute false match rate will affect the top level population estimates, while the composite false positive and negative matches may affect the accuracy of the lower level population adjustments.

6.2 The effect of an overall false match rate:

$$\gamma = 1 - \frac{\text{Number of Matches Assigned}}{\text{Number of True Matches}}$$

on a dual system estimate of the total population can be calculated explicitly within a homogeneous group of households or individuals, given the following parameters:

- the response rate to the census
- the response rate to the CCS
- the odds ratio (the probability of being counted in the CCS relative to the probability of being counted in the Census).

6.3 Table 10 shows the effect of various matching errors on the DSE of the total population for selected values of the above parameters. The figures given show the percentage error in the estimated population total that is attributable to the matching error only, and not that element that is due to the possible failure of the DSE assumption of independence between the census and CCS. This is illustrated in the two examples below.

Odds Ratio	Response Rate to Census	Response Rate to CCS	False Match Rate (1-(Matches Achieved/Total Matches))						
			0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.0%
0.1	0.98	0.90	0.1%	0.5%	1.0%	2.0%	3.1%	4.2%	5.3%
0.1	0.98	0.80	0.1%	0.5%	1.0%	2.0%	3.1%	4.2%	5.3%
0.1	0.95	0.90	0.1%	0.5%	1.0%	2.1%	3.1%	4.2%	5.3%
0.1	0.95	0.80	0.1%	0.5%	1.0%	2.1%	3.1%	4.2%	5.3%
0.1	0.90	0.90	0.1%	0.5%	1.0%	2.1%	3.1%	4.2%	5.3%
0.1	0.90	0.80	0.1%	0.5%	1.0%	2.1%	3.2%	4.3%	5.4%
1.0	-	-	-0.1%	-0.5%	-1.0%	-2.0%	-3.1%	-4.2%	-5.3%
10.0	0.98	0.90	0.1%	0.5%	1.0%	2.0%	3.1%	4.1%	5.2%
10.0	0.98	0.80	0.1%	0.5%	1.0%	2.0%	3.1%	4.1%	5.2%
10.0	0.95	0.90	0.1%	0.5%	1.0%	2.0%	3.0%	4.1%	5.2%
10.0	0.95	0.80	0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.1%
10.0	0.90	0.90	0.1%	0.5%	1.0%	2.0%	3.0%	4.0%	5.1%
10.0	0.90	0.80	0.1%	0.5%	1.0%	1.9%	2.9%	3.9%	5.0%

Table 10: Table showing the effect of matching errors on the dual system estimate of the population within a homogeneous subgroup of the population.

6.4 Care should be taken when projecting these subgroup population estimates to the total population as response rates and data quality will differ between analysis groups.

### Example 1

6.5 Consider a 98% response rate to the census and a 90% response rate to the CCS. Given a 2% matching error and an odds ratio of 0.1, the dual system estimate of the total population will be:

$$\hat{P}_{DSE} = 1.022P$$

where  $P$  is the true underlying population. However the dual system estimate in this instance, but with a perfect match would be:

$$\hat{P}_{DSE} = 1.002P .$$

Therefore the percentage error in the dual system estimate of the total population due to the matching error is 2%.

### Example 2

6.6 Consider a response rate to the census of 98% and to the CCS of 90%. Given a 2% matching error and an odds ratio of 10, the dual system estimate of the total population will be:

$$\hat{P}_{DSE} = 1.011P$$

where  $P$  is the true underlying population. However the dual system estimate in this instance, but with a perfect match would be:

$$\hat{P}_{DSE} = 0.991P.$$

Therefore the percentage error in the dual system estimate of the total population due to the matching error is 2%.

6.7 The calculations used to produce these figures are shown in Appendix 7.

## **7. Further Work**

### ***Calculation of False Match Rates***

7.1 It will never be possible to achieve 100% accuracy in the 2001 Census / CCS matching. However, it may be possible to obtain estimates of the false match rates.

7.2 One method of assessing the accuracy of the matching procedure is to review clerically a sample of matched records and to note which of these pairs appear to be true and false matches. For a reasonably accurate estimate of the false match rate to be achieved, a large clerically checked file will be required. The cost and timing of this clerical exercise may well prove prohibitive.

7.3 It may be possible to use the clerically matched data set from which the matching weights have been calculated to calibrate the system. This will only be possible if the clerical file method of allocating the probability weights is implemented.

7.4 Theoretical methods of estimating false match rates exist (Belin and Rubin (1995)), however they have been developed for matching exercises that simply match individuals and not for structured data such as those being used here. It is proposed that these theoretical methods be considered to see if they can be expanded to encompass structured data matching.

7.5 It is not clear that the theoretical estimation of false match rates will be applicable to data that have a significant proportion of hand matches.

### ***Estimation of Feasibility of Matching Exercise***

7.6 The success of the matching exercise is dependent, first and foremost, on the accuracy of the census and CCS data. It is not possible to determine the accuracy of these data in advance. However, simulated censuses and CCSs can be produced with differing error rates to assess the impact of these error rates on the matching exercise. It should then be possible to demonstrate the feasibility of the matching procedure for given levels of accuracy in the data.

### ***The Matching Programmes***

7.7 Further work needs to be undertaken in order to ensure the accuracy and efficiency of the matching programmes. It is necessary to decide on an appropriate programming language and to ensure that the resulting programmes are sufficiently user-friendly.

### ***Derivation of the Matching Weights***

7.8 It is necessary to consider in detail how the matching weights are to be derived in practice. This may involve a pre-automated matching clerical exercise to create a file of matched records from which the probability weights can be calculated. This clerical exercise

may be time consuming and costly, but the results will be intuitively and computationally straightforward. The size of the clerical file required must be calculated. It should be noted that this clerically matched file may be used to calibrate the matching process at a later stage.

7.9 Another approach may be to take the weights calculated from the dress rehearsal as starting weights and then to iteratively ‘train’ these weights within each CCS postcode, enumeration district or hard-to-count classification. The feasibility of this approach needs to be assessed.

### ***The Census Dress Rehearsal***

7.10 Once the dress rehearsal data are available for analysis it will be possible to check some aspects of the matching procedure for the first time. The contribution of address name/number and individual’s names to the accuracy of the matching procedure must be assessed. The case for the capture of these data items must then be made to justify the cost of their capture.

7.11 It will be necessary to analyse the Head of Household and Household Structure algorithms to determine their accuracy and applicability. The selection of head of household should be compared between the census and CCS to ensure that the same individual is selected in the majority of cases.

7.12 Post and pre-imputation data must be compared to see which produces the most accurate match. If post-imputation data are to be used then the variables used in donor imputation should also be taken into consideration when selecting the matching variables.

7.13 It may be necessary for the matching procedure and programmes to be reassessed after this analysis has been undertaken.

### ***Other Work***

7.14 Other issues that require consideration are:

- Housemovers
- Fuzziness of boundary edges
- Vacant properties
- String coding algorithm
- Data processing
- Asian dates of birth
- Single households in one survey being classified as multiple households in the other.
- Double counting in Census & CCS.
- Communal establishments
- Blocking within large households

## **8. Conclusions**

8.1 This paper has considered the use of probability matching for the census/CCS matching procedure. The calculation of individual matching weights has been described and implemented using prototype matching programmes. It is now possible to implement the matching algorithms on a number of simulated data sets in order to assess the impact of data errors on the accuracy of the matching procedure. This, in turn, will permit a demonstration of the feasibility of the matching exercise.

8.2 The simulated data sets can be used to test the implementation of the matching routines. However, these data sets cannot be used to test the matching procedure in full. Some areas of research cannot properly be undertaken until the data from the Census Dress Rehearsal is available for analysis. At this stage it may be necessary to revise the matching procedures.

8.3 This paper has shown that it is possible to use probability weights to distinguish between true and false matches. It has also demonstrated the implementation of the proposed matching strategy. Although many issues remain to be resolved, the basic principles of the matching have been established. Assuming a reasonable level of accuracy in the census and CCS data, this paper has demonstrated that the matching of the data sets is achievable within an acceptable time frame.

### **Appendix 1: Head of Household Algorithm**

This algorithm derives a unique representative for each household in the census and CCS datasets. This head of household's characteristics may then be borrowed as household characteristics and thus household matching variables. The algorithm attempts to select individuals irrespective of their position on the forms so that the same household representative will be chosen from census and CCS households.

1. Consider next person ( $p$ ).
2. Assign  $p$  as Head of Household ( $h$ ). Record age of Head of Household ( $h(\text{age})$ ) and whether they were working last week ( $h(\text{work}) = 1$  if working last week, 0 otherwise).
3. Consider next person ( $p$ ). Is  $p$  a member of the same household as  $h$ ? If not go to 2.
4. If  $h(\text{work}) = 0$  &  $p(\text{work}) = 1$  then go to 2.
5. If  $h(\text{work}) = 1$  &  $p(\text{work}) = 0$  then go to 3.
6. If  $h(\text{age}) > p(\text{age})$ ,  $h(\text{age}) > 65$  &  $p(\text{age}) < 66$  then go to 2.
7. If  $p(\text{age}) > h(\text{age})$  & ( $p(\text{age}) < 66$  or  $h(\text{age}) > 65$ ) then go to 2.
8. Go to 3.

## **Appendix 2: Matching Algorithms**

This section details the algorithms used to perform the automated matching.

### **1. HWt\_Calc            Household Weight Calculation Algorithm**

For a given pair of household records, compare each matching variable in turn and adjust the matching weight according to their level of agreement.

### **2. PWt\_Calc            Person Weight Calculation Algorithm**

For a given pair of individual's records, compare each matching variable in turn and adjust the matching weight according to their level of agreement.

### 3. HH\_B\_Mch( $p$ ) Within Block Matching Algorithm.

$p$  is the matching cut-off weight. Below this value no positive match will be assigned.

1. Let  $n$  be the number of CCS households within block.  
Let  $m$  be the number of Census households within block.  
Then set:  
 $i = 1.$  CCS household counter within block.  
 $j = 1.$  Census household counter within block.
2. For CCS household  $i$ :  
Set  $p_{i1}, p_{i2}$  and  $p_{i3}$  to 0. matching probabilities.  
Set  $c_{i1}, c_{i2}$  and  $c_{i3}$  to blank. census household numbers corresponding to matching probabilities.
3. For CCS household  $i$  and Census household  $j$ :
  - 3.1. Calculate matching weight using **HWt\_calc**.
  - 3.2. If weight  $< p$  goto 3.3.
    - 3.2.1. If weight  $> p_{i1}$  then:  
 $p_{i3} = p_{i2},$   $c_{i3} = c_{i2};$   
 $p_{i2} = p_{i1},$   $c_{i2} = c_{i1};$   
 $p_{i1} = \text{weight}$   $c_{i1} = \text{Census Household Number.}$
    - 3.2.2. Else if weight  $> p_{i2}$  then:  
 $p_{i3} = p_{i2},$   $c_{i3} = c_{i2};$   
 $p_{i2} = \text{weight}$   $c_{i2} = \text{Census Household Number.}$
    - 3.2.3. Else if weight  $> p_{i3}$  then:  
 $p_{i3} = \text{weight}$   $c_{i3} = \text{Census Household Number.}$
  - 3.3. If  $j < m$  then let  $j = j + 1$  and go to 3.
  - 3.4. If  $i < n$  then let  $i = i + 1$  and  $j = 1$  and go to 3.
4. Create the  $n \times 3$  matrices:  
 $P$  matching weights  
 $C$  corresponding Census household numbers.
  - 4.1. Find  $\max_{i,j} p_{ij} = p_{kl}$ , say.
  - 4.2. If  $p_{kl} = 0$  then **End**.
  - 4.3. Output CCS household identifier for household  $i$ , Census household identifier for household  $j$  and the matching weight  $p_{ij}$  to the file `h_match`.
  - 4.4. Let  $p_{kj} = 0, j = 1, 2, 3.$
  - 4.5. Where  $c_{ij} = c_{kl}$  set  $p_{ij} = 0 \forall i, j$
  - 4.6. Go to 4.1.

**End.**

#### 4. P\_Match( $p$ )      Within Matched Households Person Matching Algorithm.

$p$  is the matching cut-off weight. Below this value no positive match will be assigned.

1. Let  $n$  be the number of CCS individuals within household.  
Let  $m$  be the number of Census individuals within household.  
Then set:  
 $i = 1.$                     *CCS individual counter within household.*  
 $j = 1.$                     *Census individual counter within household.*
2. For CCS individual  $i$ :  
Set  $p_{i1}, p_{i2}, \dots, p_{im}$  to 0.                    *matching probabilities.*  
Set  $c_{i1}, c_{i2}, \dots, c_{im}$  to blank.                    *census individual numbers corresponding to matching probabilities.*
3. For CCS individual  $i$  and Census individual  $j$ :
  - 3.1. Calculate matching weight using **PWt\_calc**.
  - 3.2. If weight  $< p$  goto 3.4.
  - 3.3.  $p_{ij} = \text{weight}$                      $c_{ij} = \text{Census individual's identifier}$
  - 3.4. If  $j < m$  then let  $j = j + 1$  and go to 3.
  - 3.5. If  $i < n$  then let  $i = i + 1$  and  $j = 1$  and go to 3.
4. Create the  $n \times m$  matrices:  
 $P$                     *matching weights*  
 $C$                     *corresponding Census individual numbers.*
  - 4.1. Find  $\max_{i,j} p_{ij} = p_{kl}$ , say.
  - 4.2. If  $p_{kl} = 0$  then **End**.
  - 4.3. Output CCS individual's identifier for individual  $i$ , census individual's identifier for individual  $j$  and the matching weight  $p_{ij}$  to the file  $i\_match$ .
  - 4.4. Let  $p_{kj} = 0, j = 1, 2, 3.$
  - 4.5. Where  $c_{ij} = c_{kl}$  set  $p_{ij} = 0 \forall i, j$
  - 4.6. Go to 4.1.

**End.**

## 5. Match( $p_{h1}$ , $p_{h2}$ , $p_p$ ) Matching Algorithm

$p_{h1}$  is the household cut-off matching weight for the first block,  $p_{h2}$  is the household cut-off matching weight for the second block and  $p_p$  is the person level cut-off matching weight. Below these weights no match will be assigned.

### 1. Block on postcode and address name/number

Any census household may be linked to more than one CCS household and vice versa. Let  $B$  be the total number of distinct postcodes and address name/numbers and let  $b = 1$ .

- 1.1. For block  $b$ , call **HH\_B\_Mch**( $p_{h1}$ ). This will create a file h\_match containing matched house pairs and their matching weights.
- 1.2. If  $b < B$  then let  $b = b + 1$ , goto 1.1.
- 1.3. Create a file CCS\_um of unmatched CCS households.
- 1.4. Create a file Cen\_um of unmatched census households.

### 2. Block CCS\_um and Cen\_um on surrounding postcode and Soundex Code of HoH's Surname.

Let  $B$  be the total number of blocks and let  $b = 1$ .

- 2.1. For block  $b$ , call **HH\_B\_Mch**( $p_{h2}$ ). This will update the file h\_match to contain all CCS and census pairs matched using the two sets of blocking variables.
- 2.2. If  $b < B$  then let  $b = b + 1$ , goto 2.1.

### 3. Individual level matching within matched households

For each CCS and census household pair linked in the file h\_match:

- 3.1. Call **P\_Match**( $p_p$ ).
- 3.2. If no individual level matches are made then update the files CCS\_um and Cen\_um with the relevant CCS and census household details.

4. Recreate the file h\_match from the household details in the file i\_match so that only household pairs with linked individuals are included.

5. Create the files CCS\_um and Cen\_um of unmatched CCS and census households.

6. Clerically match the files CCS\_um and Cen\_um and the individuals within these households. Also clerically match those unmatched individuals within matched households.

**End.**

### Appendix 3: A Brief Overview of Likelihood Ratios for Matching

See “Record Linkage: Statistical Models for Matching Computer Records”, J.B. Copas & F.J. Hilton [6] Section 3 for more details.

We assume that we are attempting to match two files,  $M$  &  $N$ , on the variable  $x$ . We further assume that  $x$  can take any one of  $n$  discrete values,  $1, 2, \dots, n$ . We denote the probability of observing the value  $i$  on file  $M$  and  $j$  on file  $N$ , given that the two records belong to the same individual, by:

$$P(i, j | match) = p_{ij} \quad i, j = 1, 2, \dots, n.$$

The marginal distribution is given by:

$$p_i = \sum_j p_{ij}.$$

Then, given the hypotheses:

$H_0$ :  $i, j$  relate to different people

$H_1$ :  $i, j$  relate to the same person

the likelihood ratio of  $H_1$  versus  $H_0$  is:

$$\frac{p_{ij}}{p_i p_j}.$$

Taking natural logarithms of each estimated likelihood ratio provides evidence for and against the two records belonging to the same individual. The sign of the log-likelihood ratio indicated the direction of the evidence, positive for a match and negative against. The strength of the evidence is reflected in its absolute value.

Direct estimation of the likelihood ratios is only feasible for small values of  $n$  and with a suitably large matched file. For larger values of  $n$  another approach is required.

## Appendix 4: A Brief Overview of the Hit-Miss Model

See “Record Linkage: Statistical Models for Matching Computer Records”, J.B. Copas & F.J. Hilton [6] Section 3 for more details.

The hit-miss model is based on a binary trial, where a hit (or correct recording of a variable) occurs with probability  $1-a$ , and a miss (mis-recording of a variable) with probability  $a$ . For a miss the observed record is assumed to be distributed randomly over all possible values in proportions similar to the overall incidence of true values. Note that a miss may therefore lead to the record being correct by chance. Then the probability of observing  $i$  given that the true value of variable  $x$  is  $j$  is given by:

$$\alpha_{ij} = \begin{cases} a\beta_i & \forall i \neq j \\ 1-a(1-\beta_i) & i = j \end{cases}$$

where  $\beta_j$  is the probability that the true value of  $x$  is  $j$ . Then:

$$p_{ij} = \begin{cases} a(2-a)\beta_i\beta_j & \forall i \neq j \\ \beta_i\{1-a(2-a)(1-\beta_i)\} & i = j \end{cases}$$

$\beta_i$  is estimated as the overall relative frequency of observed value  $i$ .

When the probability of a blank,  $b$ , is sufficiently large that the probability of a double blank  $b^2$  is significantly different from zero, it is possible to take  $b$  as a random effect with mean  $\mu_b$  and variance  $\sigma_b^2$ , say.

Then:

$$p_{ij} = \begin{cases} c\beta_i\beta_j & i \neq j, 1 \leq (i, j) \leq n \\ \beta_i\{(1-\mu_b)^2 + \sigma_b^2 - c(1-\beta_i)\} & i = j, 1 \leq (i, j) \leq n \\ \{\mu_b(1-\mu_b) - \sigma_b^2\}\beta_i & j = n+1, i \leq n \\ \mu_b^2 + \sigma_b^2 & i = j = n+1 \end{cases}$$

where  $c = a(2-a-2\mu_b)$ .

The log-likelihood ratio for the pair  $(i, j)$  is then:

$$\begin{aligned} & \log c - 2 \log(1 - \mu_b) & i \neq j, 1 \leq (i, j) \leq n \\ & \log \left[ 1 + \frac{\sigma_b^2 - c(1 - \beta_i)}{(1 - \mu_b)^2} \right] - \log \beta_i & i = j, 1 \leq (i, j) \leq n \\ & \log \left[ 1 - \frac{\sigma_b^2}{\mu_b(1 - \mu_b)} \right] & i < j = n + 1 \\ & \log \left[ 1 + \frac{\sigma_b^2}{\mu_b^2} \right] & i = j = n + 1 \end{aligned}$$

where  $n+1 = \text{blank}$ .

Then estimate the required parameters as:

- $\beta_i$  relative frequency of value  $i$  amongst non-blank records
- $\mu_b$  overall relative frequency of blanks
- $\sigma_b^2$  proportion of double blanks -  $\mu_b^2$ .
- $c$  relative frequency of discordant pairs divided by  $1 - \sum \beta_i^2$ .

However, it is highly likely that the frequency of blank records may differ significantly between the census and the CCS. Therefore it may be necessary to consider two random variables, with means  $b_1$  in the census and  $b_2$  in the CCS and with common variance  $\sigma_b^2$ . The Hit-Miss model can then be expressed as follows:

$$\alpha_{ijs} = \begin{cases} a\beta_i & i \neq j; i < n+1 \\ 1 - a - b_s + a\beta_i & i = j; s = 1, 2 \\ b_s & i = n+1; s = 1, 2 \end{cases}$$

where  $s = 1$  indicates that the blank record occurred in the census and  $s = 2$  indicates that it occurred in the CCS. Then:

$$p_{ij} = \begin{cases} \{(1-b_1)(1-b_2) + \sigma_b^2 - c(1-\beta_i)\}\beta_i & i = j < n+1 \\ c\beta_i\beta_j & i \neq j; i, j < n+1 \\ \beta_i \{b_2(1-b_1) - \sigma_b^2\} & j = n+1; i < n+1 \\ \beta_i \{b_1(1-b_2) - \sigma_b^2\} & i = n+1; j < n+1 \\ b_1b_2 + \sigma_b^2 & i = j = n+1 \end{cases}$$

where  $c = a(2 - a - (b_1 + b_2))$ .

Estimate the required parameters as follows:

- $\beta_i$  relative frequency of value  $i$  amongst non-blank records
- $b_1$  relative frequency of blanks in the census
- $b_2$  relative frequency of blanks in the CCS
- $\sigma_b^2$  proportion of double blanks - (overall frequency of blanks)<sup>2</sup>.
- $c$  relative frequency of discordant pairs divided by  $1 - \sum \beta_i^2$ .

Then the log-likelihood for the pair  $(i,j)$  is given by:

$$\begin{aligned}
 & \log \left[ 1 + \frac{\sigma_b^2 - c(1 - \beta_i)}{(1 - b_1)(1 - b_2)} \right] - \log \beta_i && i = j < n + 1 \\
 & \log c - \log(1 - b_1) - \log(1 - b_2) && i \neq j; i, j < n + 1 \\
 & \log \left[ 1 - \frac{\sigma_b^2}{(1 - b_1)b_2} \right] && j = n + 1, i < n + 1 \\
 & \log \left[ 1 - \frac{\sigma_b^2}{(1 - b_2)b_1} \right] && i = n + 1, j < n + 1 \\
 & \log \left[ 1 + \frac{\sigma_b^2}{b_1 b_2} \right] && i = j = n + 1
 \end{aligned}$$

**Appendix 5: Probability Weights Calculated from the Simulated Census and CCS using Log-Likelihoods.**

This section gives the probability weights derived from the simulated census and CCS datasets using log-likelihoods.

Household level matching weights for number of people in household:

Number of People in Household		CCS						
		Blank	1	2	3	4	5	6+
Census	Blank	2.756	-5.993	-6.015	-5.408	-5.239	-4.342	-3.922
	1	-5.993	1.207	-1.593	-4.078	-6.373	-5.476	-5.403
	2	-6.015	-1.593	1.123	-1.045	-3.596	-5.296	-5.078
	3	-5.408	-4.078	-1.045	1.589	-0.067	-0.974	-4.124
	4	-5.239	-6.373	-3.596	-0.067	1.755	0.318	-1.602
	5	-4.342	-5.476	-5.296	-0.974	0.318	2.564	0.957
6+	-3.922	-4.342	-5.078	-4.124	-1.602	0.957	3.155	

Household level matching weights for Soundex code of surname of head of household:

Soundex Code of Surname of Head of Household		Census	
		Blank	Non-Blank
CCS	Blank	2.001	-0.973
	Same		5.360
	Different	-3.633	-4.119

Individual level matching weights for marital status:

Marital Status		CCS Marital Status						
		Single	Married	Re-married	Separated	Divorced	Widowed	Blank
Census	Single	0.438	-3.137	-2.259	-1.759	-2.221	-1.826	-1.263
	Married	-3.137	0.437	-1.339	-1.761	-2.223	-2.401	-2.128
	Re-Married	-2.259	-1.339	1.284	-0.883	-1.345	-1.523	-1.577
	Separated	-1.759	-1.761	-0.883	1.791	-0.345	-1.022	-1.227
	Divorced	-2.221	-2.223	-1.345	-0.345	1.348	-1.485	-1.689
	Widowed	-1.826	-2.401	-1.523	-1.022	-1.485	1.172	-1.629
	Blank	-1.263	-2.128	-1.577	-1.227	-1.689	-1.629	0.955

Individual matching weights for Soundex code of first name:

Soundex Code of First Name		Census	
		Non-Blank	Blank
CCS	Same	6.162	
	Different	-9.039	-2.565
	Blank	-2.847	2.007

Individual matching weights for Soundex code of surname:

Soundex Code of Surname		Census	
		Non-Blank	Blank
CCS	Same	6.162	
	Different	-9.039	-2.566
	Blank	-2.845	2.006

**Appendix 6: Probability Weights Calculated from the Simulated Census and CCS using the Hit-Miss Model.**

This section gives the probability weights derived from the simulated census and CCS datasets using the Hit-Miss model.

<b>Days of Birth Differ</b>	<b>Probability Weight</b>
Day of Birth differs, neither is blank	-4.848
Census Day of Birth blank, CCS not.	-0.727
CCS Day of Birth blank, Census not.	-2.208
Both Days of Birth blank	0.930

**Day of Birth the same:**

<i>Day of Birth</i>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
<i>Probability Weight</i>	1.411	1.496	1.532	1.500	1.529	1.514	1.508	1.502
<i>Day of Birth</i>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	<b>16</b>
<i>Probability Weight</i>	1.488	1.531	1.510	1.501	1.526	1.524	1.514	1.567
<i>Day of Birth</i>	<b>17</b>	<b>18</b>	<b>19</b>	<b>20</b>	<b>21</b>	<b>22</b>	<b>23</b>	<b>24</b>
<i>Probability Weight</i>	1.514	1.554	1.545	1.528	1.568	1.525	1.544	1.549
<i>Day of Birth</i>	<b>25</b>	<b>26</b>	<b>27</b>	<b>28</b>	<b>29</b>	<b>30</b>	<b>31</b>	
<i>Probability Weight</i>	1.576	1.558	1.524	1.499	1.569	1.559	1.804	

## Appendix 7: The Effect of Matching Errors on the Dual System Estimates of Population

Consider the following frequency table for individuals in a given age/sex grouping within a CCS region:

		Census	
		Included	Excluded
CCS	Included	A	C
	Excluded	B	D

Then the dual system estimate of the total population within this group (A+B+C+D) is given by:

$$\frac{(A+B)*(A+C)}{A}$$

and the odds ratio is given by:

$$\frac{A*D}{B*C}$$

Now let:  $P$  be the true population total with the given analysis group  
 $\alpha$  be the response rate to the census within this analysis group  
 $\beta$  be the response rate to the CCS within this analysis group  
 $R$  be the actual odds ratio within this analysis group.

Then we have:  $A = \frac{RBC}{D}$   $B = \alpha P - A$   
 $C = \beta P - A$   $D = (1 - \alpha - \beta)P + A$

Thus  $(R - 1)A^2 + [\alpha + \beta - 1 - R(\alpha + \beta)]PA + R\alpha\beta P^2 = 0$ .

Then the expected values of  $A$ ,  $B$ ,  $C$  and  $D$  are shown in the table below:

		Census		
		Included	Excluded	Total
CCS	Included	A	$\beta P - A$	$\beta P$
	Excluded	$\alpha P - A$	$(1 - \alpha - \beta)P + A$	$(1 - \beta)P$
	Total	$\alpha P$	$(1 - \alpha)P$	<b>P</b>

Where:

$$A = \begin{cases} \frac{P \left\{ (R - 1)(\alpha + \beta) + 1 - \sqrt{1 + (R - 1)[(2 - \alpha - \beta)(\alpha + \beta) + R(\alpha - \beta)^2]} \right\}}{2(R - 1)} & R \neq 1 \\ \alpha\beta P & R = 1 \end{cases}$$

Now, assuming a false match rate of  $\gamma\%$  of true matches we have the expected observed outcomes shown in the table below:

		<i>Census</i>		
		<b>Included</b>	<b>Excluded</b>	<b>Total</b>
<b>CS</b>	<b>Included</b>	$(1-\gamma)A$	$\beta P - (1-\gamma)A$	$\beta P$
	<b>Excluded</b>	$\alpha P - (1-\gamma)A$		
	<b>Total</b>	$\alpha P$		

Where  $A$  is as above.

This gives the DSE of the population within this analysis group as:

$$\hat{P}_{DSE} = \frac{\alpha\beta P^2}{(1-\gamma)A}.$$

## **Bibliography**

Belin, T. R., and Rubin, D. B. (1995) "A Method for Calibrating False-Match Rates in Record Linkage", *Journal of the American Statistical Association*, **Vol 90**, No. 430, Theory and Methods.

Copas, J. B., and Hilton, F. J. (1990) "Record Linkage: Statistical Models for Matching Computer Records", *J. R. Statist. Soc. A (1990)*, **153**, Part 3, pp. 287-320.

Jaro, M., (1989) "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, **Vol 84**, No. 406, Applications and Case Studies.

Kendrick, S. (1997) "The development of record linkage in Scotland: the responsive application of probability matching", *1997 Record Linkage Workshop*, 20-21 March 1997.

Kendrick, S., and Clarke, J. (1993) "The Scottish Record Linkage System", *Health Bulletin*, **Vol 51**, No. 2 March 1993.

Kendrick, S., Douglas, M., Gardner, D., Hucker, D. (1998) "Best-link matching of Scottish health data sets", *Methods of Information in Medicine*, **37**, 64-8.

Scheuren, F., and Winkler, W. E. (1993) "Regression Analysis of Data Files that are Computer Matched", *Survey Methodology*, **19**, 39-58.

**Presented at ONC Steering Committee meeting on 13 November 1998**

**Matching Process Overview**

