

ONE NUMBER CENSUS STEERING COMMITTEE

One Number Census Matching

This paper describes the work undertaken to examine the accuracy of the ONC matching process. It describes further work, undertaken on the request of the Steering Committee, following on from the preliminary analysis described in ONS(ONC(SC))00/14. The paper places the work in context of similar matching operations and examines the practicalities of performing the matching in 2001.

The accuracy of the Rehearsal clerical matching is assessed. This has led to a change in matching strategy which, coupled with improved training, will virtually eliminate the possibility of missed matches.

This paper demonstrates that, given the expected levels of data quality, the matching can be accomplished with a high degree of accuracy. We are confident of achieving a false match rate of well under the 0.1% suggested by Kendrick.

The Steering Committee are asked to note the paper.

**Jennet Woolford
Census Division
Office for National Statistics
Room 4200W
Segensworth Road
Titchfield
Fareham
HANTS
PO15 5RR**

January 2001

ONC MATCHING

1. BACKGROUND

In 2001 underenumeration will be addressed as part of the One Number Census (ONC) process. The aim of the ONC project is to produce a single census database, adjusted for the estimated undercount, so that all statistics add to 'One Number' - the national re-based estimate of the population. The major tool of the ONC will be a large, postcode based post enumeration survey, known as the Census Coverage Survey (CCS).

The data collected in the CCS will be compared with those collected in the Census to provide information about the numbers and types of households and people missed by the Census. Since the percentages missed are expected to be low (c2% of people in 1991) it is important that this comparison process, the matching, is highly accurate. Missed matches will result in an increase in the estimate of underenumeration. For example, if the true level of underenumeration is 2% and the matching process fails to identify 0.5% of true matches, underenumeration would be estimated at 2.5%.

On the advice of Steve Kendrick (ONC matching consultant, ISD, NHS Scotland) a *false negative match rate* (See Glossary) of under 0.1% should be achievable, provided the data are of the expected high quality. This paper provides further evidence for this statement.

1.1 The matching strategy

A strategy for matching the Census and CCS data using a combination of automated and clerical matching was outlined in ONS (ONC (SC)) 98/14. The key stages of the proposed matching were as follows:

1. Use blocking variables (for example postcode) for an initial grouping of the data to reduce the number of comparisons made.
2. Automatically match households within the groups defined by the blocking variables using exact and probability matching.
3. Automatically match individuals within matched household pairs using exact and probability matching.
4. Clerically review any household and individual matches where the likelihood of being a true match falls below an agreed level.
5. Clerically check any CCS households and people who remain unmatched.

The November 1998 meeting of the ONC Steering Committee approved the strategy outlined above. However it was recognised that the methodology could not be fully developed or evaluated without the appropriate data.

1.2 Evaluating the 1999 Rehearsal

The 1999 Rehearsal Census and CCS provided the first opportunity for an evaluation of the proposed matching methodology. The Rehearsal CCS was carried out in all Census Rehearsal areas in England and Wales, Scotland and Northern Ireland covering approximately 1,000 postcodes. The data were processed – captured and coded – by the Census contractor, Lockheed Martin. A full Rehearsal Evaluation Plan was drawn up which described the evaluation required to take the matching strategy forward to produce a final matching methodology (ONS (ONC (SC)) 00/06).

However, due to a combination of late delivery of data, poor data quality and the demands of the ONC evaluation timetable, the full evaluation was not possible. Thus, the evaluation has concentrated on the key issue of matching feasibility, in particular:

- the accuracy of the Rehearsal clerical matching;
- the time taken to perform the clerical matching; and
- the assistance that automatic probability matching could provide.

2. OTHER MATCHING EXERCISES

In order to put the UK work into context, the matching exercises conducted in four other census-taking countries (America, Canada, Australia and New Zealand) are considered as well as matching work conducted by the NHS in Scotland. These matching exercises are sensible, but not perfect, comparators to the UK matching. Brief descriptions of these matching exercises are given in Appendix A.

No accuracy target or achievement figures for false-negatives are available from these countries. However, the US matching system is expected to 'find and correctly link all matching records that are to be found, with a very low false match rate' (Hogan (2000)). Our matching exercise, whilst in many respects similar to the US system, differs in that we do not have the opportunity to follow up unclear matches in the field. Time constraints and the timing of the matching, particularly for the later Estimation Areas, is such that this is not feasible. As a result our matching procedures have been designed to be as accurate as possible without recourse to the field. High standards of matcher training, expert supervision and the final stage of detailed matching of unmatched households and individuals should enable us to overcome this perceived disadvantage.

The UK matching strategy requires a decision to be made about the match status of each household and individual about which data is collected in the CCS. The design of the Census form was thoroughly researched to ensure the maximum quality of the data collected. In addition, data capture techniques, interviewer and enumerator training and field procedures have all been implemented with data quality as a primary concern. Therefore, matching all CCS households and individuals is a reasonable aim.

The extensive range of NHS Scotland matching exercises more closely reflect the ONC matching strategy. These provide evidence that, given the expected high quality of the Census and CCS data, very high accuracy rates are achievable using probability and clerical matching. This is covered in more detail in section 5.

3. THE 2001 MATCHING PROCEDURE

In 2001 the Census and CCS data will be matched sequentially by Estimation Area (EA). An EA is a geographical area comprising groups of contiguous (or very close) Local Authority Districts, making up a population of approximately ½ million. The processing order for Census data has been structured so that EAs arrive in the order that most suits the matching process. Postcodes will only be matched once all contiguous postcodes have arrived and blocking will take into account these contiguous areas (i.e. EA border postcodes will be matched once the data for the neighbouring EA has arrived).

Outlined below is the updated matching strategy. The final stage is the result of the analysis of the rehearsal matching and should significantly improve the accuracy of the matching process. In addition, quality assurance checking by expert matchers has been explicitly built into stages 4, 8 and 10.

1. Use blocking variables (for example postcode) for an initial grouping of the data to reduce the number of comparisons.
2. Automatically match households within the groups defined by the blocking variables using exact and probability matching.
3. Automatically match individuals within matched household pairs using exact and probability matching.
4. Review a sample of matched households and individuals to confirm accuracy.
5. Present low probability pairs to matchers for clerical resolution.
6. Re-block unmatched data (for example by EA and phonetic coding of surname of head of household).
7. Present pairs, ranked by matching weights, to matchers for clerical resolution.
8. Expert matchers review a sample of clerically matched records.
9. Clerical search for unmatched CCS records with all blocking removed. Search at both household and individual level. Refer marginal cases to expert matchers.
10. Expert matchers review a sample of stage 9 clerical matches for quality assurance.
11. Second, independent search for unmatched CCS records remaining. Search performed by expert matchers. Search at both household and individual level.

Quality measures to ensure the accuracy of the matching will form an integral part of the process in 2001. The supervisors will ensure that standards are adhered to and difficult cases dealt with in a consistent fashion. Random samples of clerically matched pairs will be examined as matching progresses. The final stage of expert clerical matching of unmatched records will also provide an indication of the quality of the preceding clerical matching.

During the 2001 matching process probability weight thresholds will be set for the automated matching, below which records will be presented for clerical matching. These thresholds will have a direct bearing on the error rates associated with the probability matching. Threshold weights will need to be chosen such that virtually 100% of automatically linked record pairs are true matches. Theoretical methods of estimating false match rates from probability matching do exist (Belin and Rubin (1995)), however they have been developed for matching exercises that simply match individuals rather than for structured data such as those being used here. The theoretical methods will be used to provide threshold evidence and accuracy estimates during 2001. However, clerical checking of a sample of automatic matches will further ensure that if the thresholds have been set too low it will be quickly identified and adjustments can be made.

4. FACTORS IN ACHIEVING MATCHING ACCURACY

A number of key factors will contribute to the levels of matching accuracy achievable in 2001. This section outlines these factors.

4.1 Data quality

The most important factor in obtaining accurate matching is data quality. The initial proposals for the ONC matching (ONS (ONC (SC)) 98/14) were made on the unstated assumption that error rates in matching items (e.g. Soundex code of surname, year of birth, month of birth) would be of the order of magnitude of 2-3%. These error

rates are standard for administrative data such as hospital record systems, death registration etc. (Kendrick & Clarke (1993)). However, there are a number of reasons why we can expect the error rates in matching items to be better than these assumptions.

The most important of these reasons is that the data capture and processing of the Census and CCS forms will be carried out by an external contractor – Lockheed Martin. A key part of the contract is the specification of the minimum quality standards or service levels that Lockheed Martin is required to deliver. They will process the data using digital images of both tick-box and write-in data within Optical Mark and Optical Character Recognition software capable of intercepting and flagging up illegible and illogical entries. This allows extremely high accuracy to be achievable. The standards relevant for Census and CCS matching are given below:

Minimum Required Accuracy for Delivered Data

- Postcode of enumeration address 100%
- Form identifier 99.995%
- Date of Birth 99.5%
- Tick box responses 99.3%
- Alphanumeric handprint data (i.e. Name) 96%

These error rates are more stringent than the 2-3% previously assumed. In practice it is expected that the accuracy of the Date of Birth values will be higher – the process adopted is that the Date of Birth Field is captured automatically with every field being keyed from the image and the values then verified.

Of course, the data accuracy is based upon the assumption that legible, correct data are included on the actual Census and CCS forms. The CCS form is completed by interviewers and therefore control of the quality of completion is possible. Important lessons have been learnt from the Rehearsal and a much greater emphasis on data quality is being included within the 2001 interviewer-training programme.

The Census form is a self-completion questionnaire. However, the extensive work on the design of the Census form should maximise the quality of the information collected – ensuring the form is easy to understand and complete. There are also quality control mechanisms in the field and the field force will be carrying out some checks of the returned Census forms to ensure that data quality is maximised.

4.2 Accuracy of location coding

The second factor that has a large influence on the accuracy of any matching process is the accuracy of location coding. If a record's location is mis-coded then there is great potential for false negative matches (i.e. missed matches) since the true matching pair may have no chance of being compared. For example, all Bournemouth CCS records must be coded as being in Bournemouth. If a CCS record is coded with a location code (i.e. Postcode) which places it outside Bournemouth then a match may never be made.

The 2001 ONC matching strategy revolves around performing the matching area by area. Hence it is of critical importance that the location coding is accurate. This requirement is closely tied in with the requirement for a high level of data quality. As described above, the contract with Lockheed Martin specifies the levels of accuracy that must be met including the location coding. Postcode of enumeration address must be 100% accurate and the Form identifiers must be 99.995% accurate. Postcode and Estimation Area (which forms part of the form identifier) are the primary blocking variables within the matching strategy. If the stated levels of accuracy for postcode and Estimation Area are achieved then location coding should not negatively impact the false match rates. Appendix C contains a description of the rules for coding of the Postcodes for Census and CCS forms.

4.3 Flexibility of blocking criteria

In order to reduce processing time, blocking is used to restrict the number of potential matching pairs. As a result, in most matching exercises, the greatest single reason for false negative matches is inadequate flexibility of

blocking. In the ONC linkage problem, which is aimed at minimising missed matches, each CCS record must be given the widest possible range of options to find potential matches among Census records.

In order to achieve this in 2001, we will ensure that the blocking criteria are given the maximum flexibility. The final detailed search for unmatched CCS records will allow the search to cover the whole EA and bordering postcodes to ensure that the unmatched CCS record really does represent census underenumeration of an individual or household. While this may be time consuming, it is the only method of ensuring that the number of false negative matches is minimised and hence accuracy maximised.

5. COMPARISON WITH SCOTTISH LINKAGE OF HOSPITAL AND DEATH RECORDS

In order to assess the likely accuracy of the 2001 matching process it is useful to compare the ONC matching scenario with linkage exercises which have far less favourable matching conditions and for which the accuracy rates have been estimated. For this purpose we have used Scotland's standard linkage of Health and Death records. This linkage consisted of largely unstructured administrative data. The Scottish matching exercises used probability matching with only limited clerical checking to achieve false negative rates of less than 1% (Kendrick (1997)). The following highlights the enormous advantages of the ONC linkage over the Scottish linkage in terms of the factors that will have an influence on the accuracy.

5.1 Availability of population coverage target file

The target file (the Census) will have virtually 100% coverage. Employing best link probability matching methods where there is a high probability that each record in the candidate file (the CCS) will have an equivalent in the target file increases the power of the linkage immeasurably (Newcombe (1988), Kendrick, Douglas, Gardner and Hucker (1999)).

5.2 Household data – identifying information for multiple individuals

The ONC matching will link at both household and person levels. Therefore, for multi-member households, the simultaneous use of data for individual members (e.g. forename, date of birth) creates extremely rich composite household identifying information. When the proposed methodology was first developed it could not be guaranteed that name and address information would be available for analysis. The fact that this information will be available will greatly increase the accuracy of the matching.

5.3 Availability of clerical checking resources

The combination of probability matching and clerical checking can be extremely powerful. Scotland has managed without detailed clerical checking whereas the Oxford Record Linkage System (Gill (1997)) supplements its probability matching with clerical checking to improve on its levels of accuracy. The ONC strategy will use a combination of automated and clerical matching in a number of stages, which will provide significant accuracy gains over the Scottish linkage.

Therefore, given at least comparable data quality and the reasons given above, the ONC matching should achieve a false negative rate significantly lower than the 1% achieved by the Scottish linkage.

6. Evaluation Of The Rehearsal Data

The rehearsal data were matched initially using a combination of exact and clerical matching. Approximately half of the data were then independently re-matched in order to identify matching errors for further investigation. Through investigation of these errors, necessary improvements in our methodology and systems were identified which, when implemented, should enable us to meet the challenging matching accuracy targets in 2001.

6.1 Clerical matching of the Rehearsal data

The clerical matching of the rehearsal data was performed using both captured data and images. A Computer Assisted Matching System (CAMS) was developed for this purpose. CAMS performed some automatic exact matching. However, due to the strict nature of the exact matching criteria and the quality of the data, very few

complete households were matched automatically. No probability matching was used in this exercise. In 2001, with probability matching, many more records will be linked automatically and probability weights will be present to assist the clerical matcher through the presentation of likely pairs of records.

The clerical matching exercise took two weeks to complete and was undertaken by 28 different matchers, with up to six people working at any one time. The Scottish and Northern Irish data were matched by representatives from GROS and NISRA respectively. The other 26 matchers were taken from within ONS Census Division.

Table 1 below shows the headline results from the clerical matching exercise.

Table 1: Results from the clerical matching of the Rehearsal data.

Rehearsal Area	E & W	Scotland	Northern Ireland
Number of CCS Postcodes	818	130	30
Matched Households	7,681	848	168
Unmatched Census Households	1,059	148	37
Unmatched CCS Households	9,587	767	132
% Census HHs matched	88%	85%	82%
% CCS HHs matched	44%	53%	56%
Census response rate	52%	60%	59%
CCS response rate	86%	93%	85%
Within matched households:			
Matched People	14,325	1,748	352
Unmatched Census People	1,397	95	16
Unmatched CCS People	882	63	14
% Census people matched	91%	95%	96%
% CCS people matched	94%	97%	96%

The response rates shown in Table 1, especially those for the Census, are lower than those we would expect in 2001 since the rehearsal was a voluntary survey. The rates are approximate due to the difficulties of obtaining accurate figures for the total number of households from which we could expect a response. If the CCS and Census were independent, we would expect the percentage of CCS households matched to be approximately equal to the Census response rate and the Census match rate to be similar to the CCS response rate. Given the uncertainty in the response rate figures, the response and match rates shown in Table 1 are similar enough to imply a reasonable degree of independence between the data collected in the Rehearsal Census and CCS. For example, in England and Wales 88% of Census households were matched to CCS households. Since the CCS response rate is estimated at 86%, it appears that the same proportion of people responded to the CCS, regardless of whether or not they had responded to the Rehearsal Census. This implies that response to the CCS was independent of response to the Rehearsal Census.

Table 1 also shows that the Census appears to identify more people in matched households than the CCS. However, the data used for matching excluded dummy Census households, but included all proxy information collected in the CCS. Therefore there were substantially more households in the CCS that contained no individuals, largely due to households where no contact had been made and the CCS interviewer returned a household form containing proxy

household details but no person details. Amongst matched household pairs in England and Wales, 496 CCS households contained no individuals, compared with 201 Census households. If household pairs where at least one household contains no people are removed from the above figures, then the Census and CCS identified roughly the same number of people (for example 14,970 CCS individuals and 14,986 Census individuals in England and Wales). Changes to the CS interviewer training should ensure that more person information is included on proxy forms in 2001.

6.2 Re-matching the Rehearsal data

To provide us with information on matching errors a sample of approximately 50% of the rehearsal records was re-matched by ONS personnel. No records were matched and re-matched by the same member of staff. The accuracy of the clerical matching could then be assessed by examining discrepancies between the two matching exercises.

6.3 Evaluation of the re-matched data

The rematch and original match were compared to identify matching errors, explore the reasons behind them and adjust the matching strategy as necessary. The rematch revealed a total of 206 discrepancies with the original match, spread over the seven regions of England, Scotland and Wales. The discrepancies by region are presented in detail in Appendix B and are summarised, along with the measures taken to handle them, below.

6.3.1 Differences between regions

Regional differences were apparent in the matching accuracy. The rehearsal regions represent a diverse range of potential matching problems. For example, Bournemouth contains a large number of flats and holiday homes. Flats formed a large proportion of the discrepancies, but database and software problems often caused flat address information to be missed or confused by the matcher. Quality and software improvements, together with comprehensive matcher training, should enable these records to be matched accurately in 2001.

6.3.2 Matcher training

Matcher training was identified as a contributing factor in 67% (138 of 206) of the discrepancies. In the rehearsal, training was minimal, with identification of a matching pair left to the matcher's own judgement. In 2001, training and guidelines will be given to all matchers, taking into account the issues raised through the rehearsal matching exercise. Also, quality assurance measures have been put into place (see section 3). This should eliminate the majority of discrepancies between clerical matchers.

6.3.3 Data quality

Data quality was considered to be a factor in 53% (109 of 206) of the discrepancies identified. Improvements in data quality expected in 2001 are detailed in Section 4.1.

6.3.4 Variation in matching personnel

Differences were found between matching personnel. They varied both in their depth of investigation and in their judgements of the criteria determining match status. Matcher quality was considered to be a factor in 8% (17 of 206) of the discrepancies identified.

The matching of the rehearsal data was a learning exercise. Initial training was kept to a minimum and matching decisions were highly subjective. This was a particular issue where CCS forms contained address information only. In 2001, inconsistencies between matchers will be significantly reduced by more detailed training and guidelines. Supervision, as detailed in section 3, will further ensure high standards and consistency amongst matchers.

6.3.5 Software

Software deficiencies were considered to be a factor in 38% (78 of 206) of the discrepancies. The software was being trialled for the first time in the rehearsal. Necessary improvements have been identified. There is a clear need for more flexibility, allowing matchers to correct mistakes and check matches. The search capabilities and

links to address information will be improved. There will also be some changes to overall design, allowing immediate access to more information.

6.3.6 Other factors

The accuracy of the clerical matching will be strongly enhanced through the use of probability weights to rank the potential pairs of matches presented to the clerical matchers. This will remove much potential for missed matches. Automated matching will also remove the need for many records to be clerically matched, thereby reducing the potential for human error.

Marginal decisions will be referred to expert matchers/supervisors who will make the final matching decision. Remaining errors will be picked up by the final detailed match.

6.4 Accuracy Estimates for 2001

For the following reasons rehearsal results cannot be extrapolated to give quantitative accuracy estimates for the 2001 exercise:-

6.4.1 Data quality

Data quality from the rehearsal is lower than that expected in 2001. This is partly due to issues with the introduction of Optical Character Recognition (OCR), enumerator training and the rehearsal nature of the CCS. There will be considerable improvements in 2001.

6.4.2 Low coverage

Much of the power of the proposed methodology stems from the fact that the 2001 Census will approach total population coverage. Because the rehearsal was voluntary, coverage was low (c53% (Advisory Group Paper (99)09)). This is extremely important in determining the accuracy of any matching exercise (Newcombe (1988), Kendrick, Douglas, Gardner and Hucker (1999)).

6.4.3 Software and Clerical checking

The rehearsal used prototype software to assist the clerical matchers. The experiences learnt from the clerical checking of the rehearsal data will lead to improvements in the matching software, strategy and quality assurance in 2001.

6.4.4 Probability matching

Probability matching was not used to filter or to structure the presentation of pairs for clerical checking. Therefore all matching was clerical, increasing the chances of false-negative matches. Probability matching will be used in 2001.

7. TIMINGS

This section discusses the practicalities of carrying out the matching operation. Work in this area is being developed within the strategy proposed for conducting the whole of the Census downstream operation (Edit and imputation, data quality monitoring, Output – as well as ONC).

7.1 Key dates

7.1.1 Data delivery dates

The basic unit for all Census processing is the Estimation Area (EA). Lockheed Martin will deliver Census data for Estimation Areas according to a predefined order, chosen to maximise efficiency of ONC processing. The first delivery of EAs will start in **July 2001** and be completed by the end of **March 2002**.

CCS data will be processed and delivered in one block. Lockheed Martin will process CCS data as a priority and delivery is expected by **1st August 2001**.

This gives a maximum delivery window of 39 weeks for census data; allowing for one week down time for Christmas/New Year.

7.1.2 Output delivery dates

For the purposes of the Standard Spending Assessment, P&VS Division need to produce counts for Local Authorities by age and sex by **31st August 2002**.

This delivery time frame defines the timescale for ONC matching; we cannot go any faster and we cannot afford to take much longer.

7.2 Assumptions

Data arrive at (approximately) 3 to 4 Estimation Areas per week. The first delivery is likely to be at least 7 EAs. To cope with the backlog and even out the processing curve, the processing timetable is being based on 4 EAs a week.

On average:

1 Estimation Area	160 CCS Postcodes,
1 Postcode	20 Households,
1 Household	2.2 people.

The clerical matching of the Rehearsal data illustrated that the speed with which the matching is performed increases greatly with experience. However, very roughly, the Rehearsal clerical matching took approximately 10 minutes for each postcode matched.

7.3 Implementation

The basis of our current planning and input into discussions with IS for the design and operation of the Census downstream processes is the worst case scenario of carrying out a complete clerical match (with each postcode taking 10 minutes to match). However, in 2001 we would expect a great deal of matches to be assigned automatically. In addition, the probability matching facility will reveal the most likely matches to assist in the clerical matching.

We can construct alternative scenarios to examine the sensitivity of our plans. For example:

HHs requiring clerical checking	10% (true estimate believed to be closer to 3-5%)
Probability matching time per EA	10 minutes.
True population undercount	2%
Matches missed by standard matchers	1%
1 matcher to match 1 Postcode	15 minutes (10 minutes in the rehearsal)
Samples for QA checking	20%

To maintain matcher concentration an allowance of 50% of matcher time is set aside for breaks.

One day is allowed for the clerical review and subsequent setting of a threshold below which records will be passed for clerical matching.

The amount of time taken to match an EA is assumed not to be directly proportional to the number of matchers (due to likely software limitations and overlap). Therefore, based on the above figures the estimated time taken to match clerically 1 EA is as follows:-

1 matcher	1.4 days (approx.)
2 matchers	0.8 days (approx.)
3 matchers	0.6 days (approx.)
4 matchers	0.5 days (approx.)

Under the above assumptions, 3% of CCS records would require detailed clerical search. This equates to approximately 100 households per EA. In addition, the expert matchers would quality assure approximately 2% of household pairs. It is estimated that this would take two expert matchers a day to complete this work.

Under this scenario it should be possible to match the required number of EAs per week with 1 supervisor, 4 standard matchers and 2 expert matchers.

8. CONCLUSIONS

Matching is used as part of the Census taking process in many countries around the world. The closest parallels can be drawn between the ONC matching and similar exercises undertaken by NHS Scotland. Discussion with Steve Kendrick, who has extensive experience of these matching exercises, has helped form the opinion that, given the expected data quality and the increased flexibility in the revised matching strategy, there is every likelihood of achieving a false negative match rate of under 0.1%.

The re-match of the rehearsal data indicates that the initial clerical match missed 0.7% of the person matches and assigned 0.2% of matches incorrectly. Experience of performing the clerical matching will lead to improvements in the matching software and training, which will greatly reduce these error rates. In addition, the matching strategy has now been expanded to include on-going quality assessment by highly trained expert matchers. These expert matchers will also perform an additional, final search for matches, which will be unrestricted by the blocking criteria of the probability matching.

These improvements to the matching strategy, together with increased data accuracy and coverage, should reduce the matching errors to well below the 0.1% suggested by Kendrick. Therefore, provided the data is of the expected high quality, and given the timings in section 6, we are confident that **the 2001 matching can be performed within the time and accuracy constraints.**

GLOSSARY

Terminology within matching varies widely amongst the literature. The conventions adopted in this paper are as follows:

TRUE MATCH	Occurs when two linked records relate to the same entity.
FALSE NEGATIVE MATCH	Occurs when two records relating to the same entity are not linked.
FALSE POSITIVE MATCH	Occurs when two linked records do not relate to the same entity.
FALSE NEGATIVE MATCH RATE	$\frac{\text{Number of false-negative matches}}{\text{Total number of true matching pairs}}$
FALSE POSITIVE MATCH RATE	$\frac{\text{Number of false-positive matches}}{\text{Total number of true matching pairs}}$

APPENDIX A: EXAMPLES OF MATCHING IN OTHER COUNTRIES AND CONTEXTS

This annex gives a brief overview of the matching operations carried out in the context of measuring the undercount in some other countries that conduct a traditional census. In the USA, Australia and New Zealand dual system estimation is used to measure undercount, in Canada a different methodology is used.

The final example is of a large probability based matching operation carried out on National Health Service data in Scotland.

A1. USA

Measurement of undercoverage in the 2000 Census in the USA will be measured by the Accuracy and Coverage Evaluation (ACE) methodology.

The ACE sample is a stratified sample of housing block clusters containing approximately two million housing units. During September to November 1999 interviewers created lists of household units for the sampled block clusters. In January 2000 the listed sample block clusters were then sub-sampled by using a re-stratification that incorporated the ACE and updated census housing unit counts.

The sub-sampled ACE list was then matched to the Census address list— prior to Census data being available — using computer and computer assisted clerical matching, with recourse to field verification where necessary. This creates an accurate linked list of all housing units in the block.

Following the 2000 Census, the ACE interviewers attempt to attain information about all individuals living in the listed household units, interviews are conducted using laptop computers. By this stage the sample contains approximately 300,000 housing units and excludes communal establishments.

The results of these interviews are computer matched to the Census, with computer assisted clerical matching for the difficult cases. The matching is undertaken by experienced matchers, some with many years matching experience. Where necessary, field visits are conducted to determine whether a match is genuine. Any unmatched Census individuals are referred back to the field where interviewers attempt to determine whether the individual was erroneously included in the Census or was simply missed by the ACE interviewers.

The US Census does not attempt to match Census individuals for whom complete name and at least two characteristics are not available. The matching process only searches for individuals in the housing block where they believe they were enumerated or in the surrounding housing blocks in urban areas. In rural areas a slightly larger circle of housing blocks is searched and in remote areas the whole enumeration district is searched.

We could find no information about the target accuracy in the literature and contacting the US Bureau of the Census has failed to come up with anything concrete. However Hogan (2000) states that they are “confident that our matching system will find and correctly link all matching records that are to be found, with a very low false match rate.”

A2. Canada

Canada does not have a post-enumeration survey as such. Their primary coverage study is the Reverse Record Check (RRC). The purpose of the RRC is to estimate the errors in coverage of the population and of private households in the Census. It also seeks to analyse the characteristics of persons who either were not enumerated or were enumerated more than once.

The RRC uses a sample frame independent of the Census, drawing a sample of persons who should have been enumerated in the Census. A file is then created containing as much information as possible on these persons and their Census families. If possible, the addresses of the selected persons and their family members are updated

using administrative files. Retrieval operations are carried out by interviewers in order to contact the selected person and administer a questionnaire to him or her. This questionnaire determines the addresses at which the person could have been enumerated. Search operations are then carried out on the questionnaires and in the Census database in order to determine how many times the selected person was enumerated.

Probability matching is used in the address updating procedure. There are two stages to this. Firstly, probability matching is used to link the RRC file with Revenue Canada (RCT) files. Once this linkage is completed, the Social Insurance Number (SIN) of the selected person or a member of his/her family is obtained. In the second stage, an exact match is made between the RRC and more recent RCT files in order to obtain the most recent address available in those files.

Whilst probability matching the RRC sample to the first RCT file, a threshold match weight was determined for each of eight region-by-sex groups. Matches above this threshold were considered definite or possible and were retained for the next stage. The weakest links were then checked to determine their validity. This enabled an elimination of false links before proceeding and an estimation of the reliability of links retained.

The eight region-by-sex groups were matched separately. Probability weight thresholds separate out the 'possible' matches from those considered definite. Some automated checking is done on the 'possible' matches using common SIN information of spouses to inform the decision making. Remaining possible matches are then checked manually. If, after this checking, the number of rejected matches seemed high, the upper threshold is moved again to allow higher weighted matches to enter the 'possible' category. These steps were then repeated until the rejection rate for the matches checked appeared to be lower than 10% for matches with a match weight close to the upper threshold. (Bernier (1997))

Canada also performs an automated match study (AMS) to detect over-coverage that occurs between private dwellings. Central to the AMS are a series of computer programs, which identify pairs of households that are 'similar'. Similarity is described in terms of the sizes of the two households, their relative geographic proximity and the number of person matches between them. Since names are unavailable for matching, persons are matched on the basis of sex and date of birth. Two persons with the same sex, day, month and year of birth are said to exactly match. If three of the four components agree, or if just the day and month of birth are transposed, persons are said to nearly match. The pairs of households identified constitute the survey's sampling units and were stratified by province and similarity. The Census questionnaires for a sample of pairs from each strata were verified to determine how much over-coverage, if any, was present. This determination was made on the basis of the names found on the questionnaires. Estimates of total over-coverage were then produced for a variety of sub-populations by weighting up the survey data. No accuracy figures are available for the AMS matching exercise (Ha, Mayda & Tourigny (1998)).

A3. Australia

In Australia a Post-Enumeration Survey (PES), conducted independently of the Census is used with DSE to estimate under-count (Dunstan et al (1999)). A multi-stage stratified sample is drawn from the ABS Labour Force Survey, sampling approximately 0.5% of private dwellings. The PES collects information via a face-to-face interview asking respondents for an address where they may have been included on a census form. Visitors are included in the PES and asked for their address of usual residence. In 1996 there were 31,200 fully responding households.

The PES collects name, sex and date of birth or age to facilitate accurate matching. Also collected are marital status, country of birth and indigenous origin – these can be used for more difficult matches.

Clerical matching of individuals was performed using the physical Census and PES forms. Visitors who may have been enumerated in non-sampled households were matched back to the addresses where they felt that they may have been enumerated. These responses were used to determine the number of times that each respondent was included in the Census.

Where address information was too vague to allow match status to be determined, match status was imputed (in 1996 1,600 of the 7,300 search addresses - about 22%). The value imputed onto the PES record was an integer reflecting the number of times the individual was captured in the Census (0 - corresponds to missed from the Census, 1 - counted in the Census precisely once, 2 - counted in the Census twice etc.)

A4. New Zealand

The New Zealand procedures for measuring and adjusting for undercount are similar to those adopted in Australia. The NZ PES is a two-stage stratified cluster sample household units drawn from the SNZ Household Labour Force Survey sample frame (Dunstan et al (1999)). The sample covers approximately 0.8% of total private dwellings in NZ. The information collected in the PES is similar to that for the Australian PES, detailed above.

Matching of individuals is a clerical exercise, using physical PES forms and the images of Census forms. Where address information is insufficient, the match status was imputed.

Both Australia and NZ plan to use similar matching strategies in 2001 to those outlined above.

A5. National Health Service in Scotland

The National Health Service (NHS) in Scotland has performed over 150 separate matching exercises in the 5 years up to 1997. Primarily, these have involved linking external data sets (e.g. Survey data, clinical audit) to the centrally held health records. The exercises have varied enormously in scale and complexity. In the creation of the national linked data sets, probability matching combined with only a limited amount of clerical checking, resulted in false - negative rates of less than 1% (Kendrick (1997)) Probably the largest and most relevant work to the ONC was the linkage of the Community Health Index (CHI) and the National Health Service Central Register (NHSCR) in Scotland. This linkage is now described in more detail.

The linkage combined deterministic and probability matching techniques. Comparison of CHI records with NHSCR records was done in three passes through the data. Each pass brought together pairs of records sharing the same 'blocking criteria'. Pass 1 brought together records sharing the same 12-character NHS number and used both deterministic and probabilistic matching. Passes 2 and 3 performed probability matching using different sets of blocking criteria. Results were tracked for each CHI record by storing the highest link weight achieved so far with the unique identifier of the corresponding NHSCR number.

Surname, forename, gender, date of birth, previous surname, Health Board and date of death (where relevant) were available on both Scottish files. The CHI records contained "date of acceptance by GP practice" which could be compared with "date of transfer to current Health Board" on the NHSCR record. NHS numbers were available on all the NHSCR records and the majority of the CHI records.

There are important similarities between the data for this Scottish linkage and the proposed UK matching in 2001. In both cases the percentage coverage in the file to be matched is very high. Both exercises use extremely large numbers of records (the CHI and NHSCR files covered almost the whole population of Scotland - over 5 million records). In Scotland, there should be only a single NHSCR record for an individual and in the UK Census there should only be one record for an individual in either the CCS file or the Census file. Although the Scottish matching has the apparent advantage of a unique identifier in the NHS number, the NHS numbers in the CHI are of limited use to the matching because of the wide variety of formats and recording difficulties. The Census has some similar person-identifying information to the NHS (Surname, gender and date of birth) but also has address and residency details, providing richer information for matching.

Conversion from relative to absolute odds of a record being a true match can be improved by retaining only the best (highest weight) link achieved. Because the Scottish residents were likely to be represented by one NHSCR record and one or more CHI records, the linkage could be structured as a best-link many-to-one linkage (Kendrick,

Douglas, Gardner and Hucker (1999)). Since, in 2001, the CCS and Census files should each only contain a single record for an individual, we should also be able to employ a best-link linkage.

Accuracy

A 'best-estimate' of the number of linkable CHI records was made by determining by clerical inspection the 50:50 threshold (i.e. the weight at which it is equally likely that the 2 records belong or do not belong to the same person) for best links. Operationally however, the important threshold is that above which the administrators responsible for operating the new system are willing to accept links as being sufficiently accurate for administrative purposes. To this end, the Primary Care teams in each of the Scottish Health Boards checked a sample of pairs across the entire weight range to determine this threshold. No Health board required a threshold of >30 (the 50:50 threshold was 15). Of 500 pairs with $\text{weight} > 30$ which were fed back after checking, 0 were found to be incorrect. Subsequent checking by NHSCR of 2000 pairs with a weight greater than 30 failed to find an incorrect link. 98.8% of the CHI records estimated to be linkable by a perfect linkage were linked either deterministically or with a probability weight of 30 or more.

APPENDIX B: DESCRIPTION OF DISCREPANCIES FOUND IN THE 50% REMATCH

B1. Household Discrepancies

Lincoln

There were 51 discrepancies (1381 true matches found). Of these, 46 should be eliminated in 2001 by a combination of improvements in matcher training and/or data quality (quality improvements alone will eliminate 9 of these 46 discrepancies). Matcher quality was identified as the primary factor in a further 4 discrepancies. The 1 remaining discrepancy is ascribed to a marginal decision. The final stage of detailed clerical matching and supervisor support should help to eliminate this final error.

Bournemouth

There were 51 discrepancies (619 true matches found). Of these, 41 should be eliminated in 2001 by a combination of improvements in matcher training and/or data quality. A combination of quality and software improvements will eliminate a further 5 discrepancies. The remaining 5 discrepancies are ascribed to marginal decision. The final stage of detailed clerical matching and supervisor support should help to eliminate these final 5 errors. Approximately 47 of the 51 discrepancies referred to flats. There is no doubt that most of the discrepancies were made difficult to match by an absence of address information presented to the matcher. Although matcher training would have prevented the vast majority of discrepancies, software improvements and/or data quality improvements would have assisted in 39 of the 51 cases.

Ceredigion

There were 16 discrepancies (321 true matches found). Of these, 12 would be eliminated by a combination of improvements in matcher training and/or data quality. A combination of software and quality improvements will eliminate a further 2 discrepancies. The remaining 2 discrepancies are ascribed to marginal decision. The final stage of detailed clerical matching and supervisor support should help to eliminate these final 2 errors.

Gwynedd

There were no discrepancies in Gwynedd (133 true matches found).

Dundee

The Census addresses were not automatically displayed in the matching software. This was due to problems specific to the rehearsal regarding links to the geography database and will not occur in 2001. This absence of Census address made searches for matches very difficult. This may well have been the primary reason for most of the discrepancies. There were a total of 16 discrepancies (104 true matches found). Of these 14 would be eliminated in 2001 by a combination of improvements in matcher training and/or data quality, and software correction. The remaining 2 discrepancies are ascribed to matcher quality and marginal decision. The final stage of detailed clerical matching and supervisor support should help to eliminate these final 2 errors.

Angus

The Census addresses were not automatically displayed in the matching software. This was due to problems specific to the rehearsal regarding links to the geography database and will not occur in 2001. This absence of Census address made searches for matches very difficult. This may well have been the primary reason for most of the discrepancies. There were a total of 15 discrepancies (326 true matches found). Of these 14 would be eliminated in 2001 by a combination of improvements in matcher training and/or data quality, and software correction. The remaining 2 discrepancies are ascribed to matcher quality and marginal decision. The final stage of detailed clerical matching and supervisor support should help to eliminate these final 2 errors.

Leeds

There were 57 discrepancies (1070 true matches found). Of these, 43 should be eliminated in 2001 by a combination of improvements in matcher training and/or data quality. A combination of quality and software improvements will eliminate a further 7 discrepancies. The remaining 7 discrepancies are ascribed to matcher quality and marginal decision. The final stage of detailed clerical matching and supervisor support should help to

eliminate these final 7 errors. Although matcher training would have prevented the vast majority of discrepancies, software improvements and/or data quality improvements would have assisted in 31 of the 57 cases.

B2. Summary Of Household Discrepancies By Type And Area

Codes

1. Poor data capture (data quality)
2. Poor Enumeration (data quality)
3. Poor matcher training
4. Poor matcher quality
5. Software deficiencies
6. Poor/confusing information (CCS form)
7. Poor/confusing information (Census form)
8. Marginal decision

Clerical Matching Discrepancies – Households						
	Lincoln	Bournemouth	Ceredigion	Dundee	Angus	Leeds
1	1	1	5			
2	8	1	1		1	1
3	22	2	3			12
4	3					1
5	1		1	3	5	3
6						
7						
8		5	2			2
1,2						1
1,3	5	6	3			2
1,5		4	1	3		1
2,3	3	1				2
2,5		1		1		2
2,8						2
3,4	3	1				7
3,5	1	3		2	6	4
3,7	2	2				2
4,7	1					
5,7				1	1	1
5,8				1		1
1,2,3		2				6
1,3,4						1
1,3,5		13		4	1	3
1,3,6	1					
2,3,5		7				
2,3,8						1
3,5,8					1	
3,6,7		2				
5,7,8						1
1,2,3,5				1		
2,3,5,6						1
Total	51	51	16	16	15	57

B3. Person Discrepancies

The majority of discrepancies between the matching of person are due to unclear ground rules for the matchers. In 2001, Census individuals will only be created if a certain level of person information is available. In the rehearsal, different matchers applied different rules. Rules will be devised in 2001 for the matchers when dealing with person records in the CCS.

Lincoln

There were 13 discrepancies contained within a total of 9 households (total true person matches = 2878). Matcher guidelines and training should eliminate 4 of these discrepancies. Poor information on the two forms caused the confusion on 7 of the discrepancies, and the matching decision was therefore marginal; matcher training and reference to a supervisor would overcome this problem (the final detailed clerical match would provide a further safety net). One discrepancy could not be explained except by careless matcher error. Software correction would eliminate the final discrepancy.

Bournemouth

There were 5 discrepancies contained within a total of 5 households (total true person matches = 1053). Matcher training should eliminate 3 of the discrepancies. Matcher training may eliminate one more discrepancy but it is primarily a matcher quality issue with insufficient care taken when matching the last individual of a household. The final discrepancy was a marginal decision and the final stage of detailed clerical matching coupled with supervisor support should eliminate this error.

Ceredigion

There were 4 discrepancies contained within a total of 3 households (total true person matches = 672). 3 of the discrepancies involved differences in the forenames between Census and CCS and should be eliminated by matcher training. The remaining discrepancy required reference to the image and should also be eliminated by matcher training.

Gwynedd

There were 0 discrepancies (total true person matches = 319).

Dundee

There was 1 discrepancy (total true person matches = 218). The records contained different information on Christian names and there were some capture problems. The discrepancy should be eliminated by matcher training.

Angus

There were 3 discrepancies contained within a total of 2 households (total true person matches = 668). 1 discrepancy will be eliminated through matcher training. The other 2 were a father and son with identical names and data capture problems on date-of-birth. Matcher training should help to eliminate these errors.

Leeds

There were 37 discrepancies contained within a total of 27 households (total true person matches = 1759). The pattern of causes for these discrepancies is more complex than those in the other regions. The discrepancies generally arise from a combination of causes, most of which will be eliminated in 2001 as described above. However, there was more of a problem with quality of matching personnel for this region, with one matcher in particular seeming to have concentration problems. Matcher quality was an issue in 11 of the discrepancies. Matcher training, combined with the improvements in data quality and software for 2001 should eliminate a further 24 discrepancies. The remaining two discrepancies were marginal cases and matcher training in 2001 should guide the matchers to refer the cases to a supervisor (the final detailed clerical match would provide a further safety net).

B4. Summary of Person Discrepancies by Discrepancy Type and Area

Clerical Matching Discrepancies – Person Records						
	Lincoln	Bournemouth	Ceredigion	Dundee	Angus	Leeds
3	2	2	1		1	5
4	1					8
5	1					
8		1				2
1,2						5
1,3	1					1
1,4					2	
3,4		1				1
3,7			3			1
4,7						2
1,2,3						1
1,3,6				1		
2,3,5						2
2,3,6						1
3,4,5	1					
3,6,7	2					5
3,4,6,7		1				
3,5,6,7						2
3,6,7,8	3					
4,6,7,8	2					
1,2,3,6,7						1
Total	13	5	4	1	3	37

APPENDIX C: PROCESSING RULES FOR POSTCODES

The capture and processing of Census and CCS forms is carried out by an external contractor - Lockheed Martin. Outlined below are the requirements for the processing and coding of postcodes.

C1. Census

Fieldwork

The Enumerators Record Book (ERB) provides the Enumerators with list of the addresses that exist on Address Point for their ED. Each of the addresses in the ERB is allocated a form number that the Enumerator transcribes onto the Census form when it is delivered.

However Address point is not complete so there are addresses that the enumerator finds that are not pre-printed. In this case the enumerator adds the address (including the postcode to the ERB).

Forms are posted back to the Census District Manager or returned to him/her via the follow-up. They are ultimately delivered to LM in boxes of whole EDs, sorted by form numbers and accompanied by the appropriate ERB.

Any Census forms that come back after this are dealt with centrally.

Processing

1. Derive postcode from the Form ID

The Form ID consists of CD (Census District), ED (Enumeration District) codes and a four-digit Form number. It is a critical field and is captured automatically and also verified by keying from the image.

The Form ID is checked for validity. Where it is wholly valid it is matched against the Geography database to derive a postcode. The majority of cases (95%) will be matched in this way.

If any of the elements of the ID are missing or invalid, all available relevant information is used to obtain the correct value e.g. adjacent forms and boxes, the ERB. If it cannot be established then the issue is referred back to the Authority (Census office).

2. Enumeration Postcode

If the postcode is not derived from the Geography Database, then the address and postcode are captured from the forms and the last two digits of the postcode and relevant parts of the Form ID are used to match against the Geography database to get a full postcode.

3. Enumeration Address

If there is no postcode, or it is not matched on the Geography database, or it is invalid for the ED and there is an address present, then use the address to obtain a postcode by matching the address against a postcoding package and verify the postcode against the geography database and the ED.

If the postcode is not found from the address then when all possible sources of information have been checked and the postcode still cannot be found, then refer to the Authority.

C2. Census Coverage Survey

Fieldwork

The CCS is an independent re-enumeration of selected postcodes. Thus it is not based on an address list. The interviewers are given maps of their areas and required to ensure that they identify all households with their allocated postcode sample.

The correct identification of all households with the selected postcodes and the correct recording of the postcode on the CCS forms is crucial to the success of the CCS and the training of the field staff reflects this.

Processing

Capture Postcode and Address

a) Postcode Present

The Geography Database contains a list of the postcodes included in the CCS.

Postcodes are validated against the Geography Database and checked against the Estimation Area (EA) derived from the Form ID to check that the postcode is in the correct EA.

(The CCS Form ID consists of Interviewer Number and Form Number. The Interviewer Number consists of Country code, Team Manager Code, CCS Workload and Interviewer code. It is a critical field and is captured automatically and also verified by keying from the image).

If it is not on the CCS Postcode list or is inconsistent with the EA then a postcode is generated from the address using a postcoding package. This is then validated against the list of CCS postcodes and the EA.

If it is not on the list or not in the correct EA then it is referred to Authority who will use all available information to provide a valid Postcode (e.g. use CCS equivalent of ERB, use management information systems)

b) Postcode Missing

If possible, generate a postcode from the address and validate against the list of CCS postcodes and the EA. Otherwise refer to Authority.

c) No Postcode or Address

Refer to Authority.

REFERENCES

- Baxter, J. (1998) "One Number Census Matching", ONS(ONC(SC))98/14.
- Bernier, J. (1997) "Quantitative Evaluation of the Linkage Operations of the 1996 Census Reverse Record Check", Federal Committee of Statistical Methodology, Proceedings of an International Workshop and Exposition 'Record Linkage Techniques' Chapt.5. p160-167.
- CPB(00)18. Downstream Processing Strategy for the 2001 Census.
- CPB(00)41. Update on Operational Timetable.
- Dunstan, K., Heyen, G. and Paice, J. (1999) "Measuring Census Undercount in Australia and New Zealand", Australian Bureau of Statistics, Demography Working Paper No. 99/4
- Gill, L. E. (1997) "OX-LINK: The Oxford Medical Record Linkage System", Federal Committee of Statistical Methodology, Proceedings of an International Workshop and Exposition 'Record Linkage Techniques' Chapt.2. p15-33.
- Ha, B., Mayda, M. and Tourigny, J. (1998) "Methodology of the 1996 Automated Match Study", (draft) Statistics Canada, Internal Report.
- Heady, P., Smith, S. and Avery, V. (1994) "1991 Census Validation Survey: Coverage Report", OPCS, London HMSO.
- Hogan, H., (2000) "The Accuracy and Coverage Evaluation: Theory and Application" Prepared Paper for The National Academy of Science DSE workshop. 2-3 February 2000.
- Kendrick, S., (1997) "The Development of Record Linkage in Scotland: The responsive Application of Probability Matching". Federal Committee of Statistical Methodology, Proceedings of an International Workshop and Exposition 'Record Linkage Techniques' Chapt.10. p319-332.
- Kendrick, S. and Clarke, J. (1993) "The Scottish Record Linkage System", Health Bulletin (Edinburgh), Vol 51 No.2 March 1993.
- Kendrick, S. W., Douglas, M. M., Gardner, D. and Hucker, D. "Best-Link Matching of Scottish Health Data Sets", Yearbook of Medical Informatics 1999, p405-409.
- Newcombe, H. B. (1988) "Handbook of Record Linkage", NY: Oxford University Press 1988.
- ONS(ONC(SC))00/06. Rehearsal Evaluation Plan
- ONS(ONC(SC))00/14. ONC Matching
- ONS AG Paper (99)09. "1999 Census Rehearsal Update"