

ONE NUMBER CENSUS STEERING COMMITTEE**ONC Matching**

1. This paper demonstrates that the 2001 ONC Matching is achievable within the necessary time and accuracy constraints. It also shows that automatic probability matching can make a significant contribution to the matching process.
2. **Members of the board are asked to approve the decision to perform probability matching followed by clerical matching in 2001.**

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

June 2000

ONC Matching

1. Summary

This paper provides an update on the research on matching. The previous matching methodology paper ONS(ONC(SC))98/14 describes the proposed methodology as far as it could be developed without data for analysis. The Rehearsal Evaluation Plan (ONS(ONC(SC))00/06) describes the evaluation required to take the matching strategy forward to produce a final matching methodology.

Due to the lateness of the delivery of the Rehearsal data it has not been possible to evaluate the probability matching methodology fully in time for inclusion in this paper. This paper therefore addresses the key issue of matching feasibility.

This paper considers:

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

The paper demonstrates that the 2001 matching exercise can be performed within the necessary time and accuracy constraints. It argues further that automatic probability matching can make a significant contribution to the timeliness and accuracy of the overall matching.

It is therefore recommended that the matching of the CCS and Census data is undertaken using a combination of automatic and clerical matching in 2001.

2. Background

In the 1991 UK Censuses, the estimation of underenumeration was made problematic by its differential nature. 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 rebased 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).

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 households and individuals who do not appear on a census form. The accuracy of the matching is of critical importance. Even a few matching errors may have a sizeable effect on population adjustments. In his letter of 7 November 1998, Steve Kendrick states that the matching exercise should aim for a false negative (missed matches) rate for individuals of between 0.02% and 0.1%.

A combination of automated and clerical matching is proposed. The automated matching process has been structured to utilise the hierarchical nature of the census data. The overall matching procedure allows for clerical checking of those linked pairs that have a low probability matching weight, as well as a clerical search for matching census records for all unmatched CCS households and individuals.

The key stages of the proposed matching, as described in ONS(ONC(SC))98/14, are as follows:

1. Use blocking variables (e.g. 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 with 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.

Whilst other areas of ONC methodology have been tested using simulated data, this was not feasible for the matching methodology. Therefore, the 1999 Census Rehearsal data provide us with the first opportunity to assess the proposed matching strategy.

3. Clerically Matching the Rehearsal Data

The first stage of the ONC matching Rehearsal evaluation involves matching clerically the Rehearsal CCS and Census data to establish the true pairs of households and individuals.

The Rehearsal matching 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.

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.

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

Overall, the performance of the Computer Assisted Matching System was very successful. Feedback from the clerical matchers was positive and although many had suggestions for how the system could be improved, these suggestions were reasonably trivial to implement.

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

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%

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

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.

It can further be seen from Table 1 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).

4. Accuracy of the Clerically Matched Rehearsal Data

Since clerical matching is the final stage of the matching in 2001 it is necessary to ensure that the quality of the clerical matching is sufficient for our purposes.

Two measures of accuracy are commonly used when performing matching:

- ***False positive matches*** occur when two linked records do not relate to the same entity.
- ***False negative matches*** occur when two records relating to the same entity are not linked.

The false match rates are then defined as:

$$\text{False Positive Match Rate} = \frac{\text{Number of false positive matches}}{\text{Total number of true matching pairs}}$$

$$\text{False Negative Match Rate} = \frac{\text{Number of false negative matches}}{\text{Total number of true matching pairs}}$$

The accuracy of the clerically matched data was assessed in two ways.

- 48 Rehearsal postcodes, selected randomly from Bournemouth, Lincoln and Gwynedd, were re-matched using CAMS. The results from the two matching exercises were compared and those households and individuals matched by just one of them were investigated.
- All unmatched Census households in the Bournemouth CCS areas were considered in detail to determine whether a matching CCS household could be identified.

4.1 Re-matched postcodes

An analysis of the 48 postcodes re-matched following the clerical matching exercise gave the results shown in Table 2 below.

HOUSEHOLDS	
Number of true matching pairs:	410
Number of false positives:	1
Number of false negatives:	6
False positive match rate:	0.2%
False negative match rate:	1.5%
PEOPLE	
Number of true matching pairs:	893
Number of false positives:	1
Number of false negatives within unmatched households:	7
Number of false negatives within matched households:	1
False positive rate:	0.1%
False negative rate:	0.9%

Table 2: Estimated accuracy of the Rehearsal clerical matching exercise.

It is clear from Table 2 that once household pairs have been identified the clerical matching of individuals within household pairs is highly accurate. Therefore the matching software and training in 2001 should focus on the household level matching.

It is important to note that it is not possible to project the accuracy of the matching in 2001 from the accuracy of the Rehearsal matching. However, the error rates demonstrated here are very low once consideration is taken of the issues of data quality, response rates and the experimental nature of the clerical matching training and software. This is a powerful indication that the challenging accuracy rates required for 2001 will be achievable.

4.2 Unmatched Bournemouth Households

CAMS only offered users the opportunity to match households in the same or neighbouring postcodes. The unmatched Bournemouth households were considered in detail, and an extensive search was performed using different variables from those used by the matching software.

The more extensive search here included:

- consideration of the images of the forms to ensure that the address had been captured correctly.
- searching using the names of household members.
- searching for addresses outside the local set of postcodes.

This analysis found 19 missed household matches out of 1,298 potential matches. This gives a household false negative match rate of 1.5%, consistent with the findings from the re-matched postcodes.

This exercise provided valuable insight into how the final stages of the clerical matching process can be improved for live running in 2001.

5. Probability Matching

In order to demonstrate the contribution that automatic probability matching can make to the overall matching process, an initial probability match was undertaken using matched Rehearsal data. This matching was performed at the household level only.

Simple matching weights were calculated from the matched Bournemouth data for address, tenure, household type and an encryption of the surname of the head of household (see ONS(ONC(SC))98/14 for more details on calculating probability weights). Matched Leeds data was then used to simulate a Census and CCS with response rates of 95% and 90% respectively. Leeds data was used in an attempt to simulate using weights from one set of data to match data from another source. However, this will affect the quality of the probability matching as the characteristics of the Leeds CCS areas are very different from those in Bournemouth. The simulated Census and CCS data contained 2,069 households in common.

A very simple matching programme linked the simulated Census and CCS data using the Bournemouth weights. This programme automatically linked 1,953 pairs of households. 1,905 were true matching household pairs and 49 were false pairs. The distribution of weights for true and false matching pairs is shown in the Figure 1 below. The absolute values of these weights are unimportant, they merely serve to rank possible pairs of records with respect to the likelihood that they related to the same household.

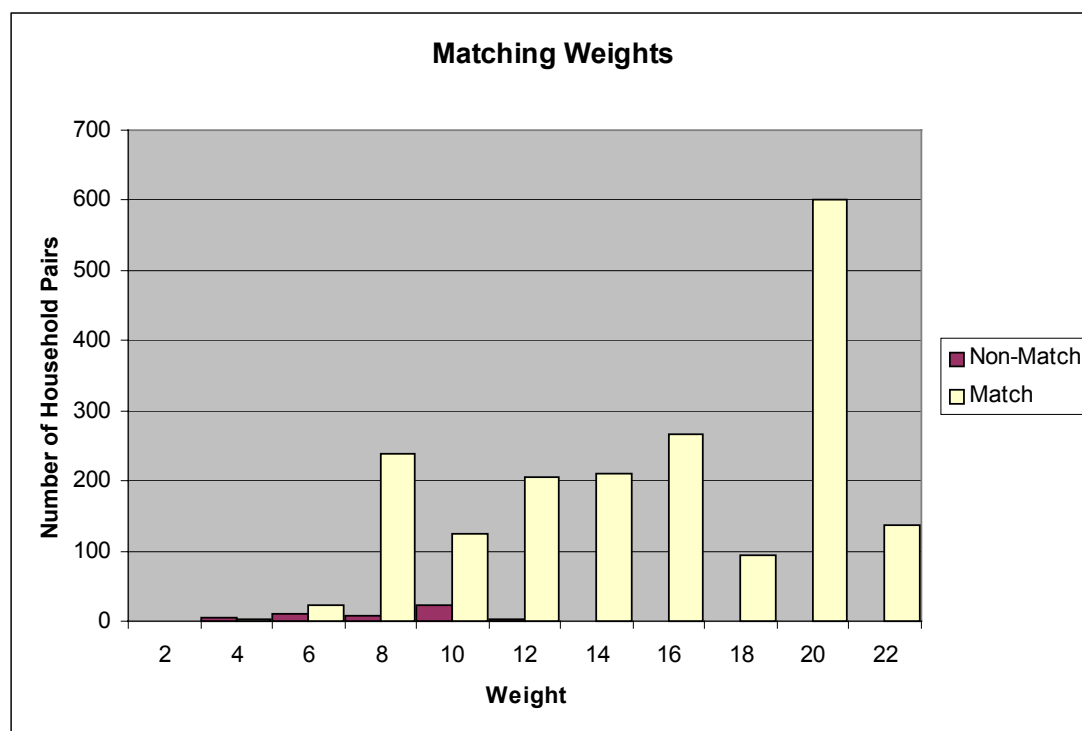


Figure 1: Graph showing the probability matching weights for true matching and non-matching household pairs.

It can be seen from Figure 1 that it is possible to impose a cut-off weight that would eliminate all or most of the spurious pairs of households. For example, if pairs were only matched where they had a weight of 11 or over, 1,328 household pairs would be identified, of which one would be a spurious match (false positive rate of 0.08%). This leaves approximately 40% of CCS households for clerical consideration.

A more sophisticated matching programme and set of matching weights and variables should be capable of more accurate matching than the preliminary results given here. However, this example clearly illustrates that probability matching can be used reliably to reduce the amount of clerical matching required.

6. Conclusions

6.1 Accuracy in 2001

Clerical matching accuracy should be better in 2001 than that experienced in the Rehearsal due to the following:

- In Kendrick's letter of 7 November 1998, he lists seven aspects of the proposed linkage which support the belief that acceptable accuracy should be achievable. The first of these is the high expectation (approximately 98%) that an individual in the CCS will be represented in the Census. This expectation does not hold in the Census rehearsal, where the Census response rate for England and Wales is estimated at 52%.
- Lessons learned from the Rehearsal matching will lead to improvements in the matching software and training.

- The quality and quantity of the information on the Census and CCS forms is expected to be higher in 2001. This will make identifying matches more straightforward.
- The use of probability matching will reduce the percentage of matching records that have to be clerically identified.
- The Rehearsal CCS areas were purposefully selected to allow the investigation of specific enumeration problems such as students, multi-occupancy and ethnic mix. Therefore results observed in these areas would be expected to be less accurate than those for the nation as a whole.

Considering the above points and the encouragingly low false match rates identified in the Census Rehearsal, it is concluded that **the matching can be performed to the required levels of accuracy in 2001.**

6.2 Timing in 2001

Data is expected to be delivered at a rate of approximately four Estimation Areas a week in 2001. On average, each Estimation Area will contain around 200 CCS postcodes. It is therefore estimated that, on average, it will take one person around one week to match each estimation area. This timing sits comfortably within the current timetable and staffing projections. Therefore, **the matching should be possible within the time constraints.**

6.3 Probability Matching

Further evaluation of the Rehearsal matched data will allow the improvement of the automatic matching to achieve more matches automatically. Also, expected improvements in the response rates and the quality of the data in 2001 will lead to more automatic matches being achieved. However, it is clear from these preliminary results that **probability matching can make a significant contribution to the matching exercise in 2001.**