

ONE NUMBER CENSUS STEERING COMMITTEE**Matching Error**

1. This paper outlines the work that has been undertaken to measure and improve the accuracy of the ONC matching process. This work has successfully achieved a false match rate significantly below the 0.1% suggested by Kendrick (ONS(ONC(SC))00/14).
2. The paper details the steps that have been taken to improve accuracy and the effect that they have had. It also explains the steps that are being taken to further improve accuracy.

The Steering Committee are asked to note the paper.

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

February 2002

Matching Error

1. Summary

- 1.1 This paper outlines the work that has been undertaken to measure and improve the accuracy of the ONC matching process. This work has successfully achieved a false match rate significantly below the 0.1% suggested by Kendrick (ONS(ONC(SC))00/14).

2. Background

- 2.1 The One Number Census (ONC) project aims to estimate underenumeration in the 2001 Census and produce a single census database adjusted for this estimate. In order to produce estimates of underenumeration a large postcode based post enumeration survey, known as the Census Coverage Survey (CCS), was carried out. The results of the Census and the CCS are compared in order to provide estimates of the numbers and types of people and households missed by the Census.
- 2.2 For the estimates of underenumeration to be accurate it is important that the comparison process, or matching, is highly accurate. Inaccuracies in the matching process have a corresponding effect on the ONC estimates. If the matching process failed to identify 0.5% of true matches, then underenumeration would be estimated at 0.5% higher than the true level.

3. Measuring matching accuracy

- 3.1 To establish the accuracy of the matching process and ensure that the output is correct prior to ONC estimation, all Design Groups have been matched several times. The output from each matching exercise is compared, and inconsistencies are investigated. As a result of these investigations, a final decision is made on which version of the output is correct, and which represents an error in each instance.
- 3.2 There are two types of matching error:
 - False negative errors; two records relating to the same entity are not linked, or the match is missed
 - False positive errors; two records are linked although they relate to two different entities.
- 3.3 False positive matches are extremely rare in this matching exercise. The high quality of the geographical, household and person information available means that it is highly unlikely that a CCS individual will be coincidentally either:
 - more similar to a non-matching Census record than to the true matching Census record; or
 - not have a true matching Census record and be similar enough to a non-matching Census record (which must also not match to any other CCS record) for a match to be assigned.

3.4 This paper concentrates on false negative matches. The false negative match rate is defined as follows:

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

3.5 Errors from the matching of early Design Groups were examined to establish how they occurred and might be avoided in future. Changes in the matching strategy were then considered in order to ensure that errors of this type did not reoccur. Once all errors have been investigated, the estimation area is matched for a final time, ensuring that the final output contains none of the errors identified.

3.6 In order to measure matching accuracy each Design Group must be matched at least three times, twice for the comparison and once to produce the final version of output for estimation. This method has ensured that the output that goes to ONC estimation contains no errors.

4. False match rates

4.1 **Table 1** below shows the re-matched results for three Design Groups. The figures shown relate to person matches. It should be noted that all identified errors are corrected prior to estimation and therefore the ultimate false negative match rates for each of these areas is an estimated 0.00%.

Table 1: False match rates identified in the first match of three early Design Groups. All errors are corrected before ONC estimation takes place.

	Design Group 1	Design Group 2	Design Group 3
Number of true and false matches	4756	5652	5359
False negative matches identified in first match of area	63	16	3
False negative match rate identified in first match of area (%)	1.32	0.28	0.06

4.2 The Design Groups in **Table 1** are shown in chronological order and a clear improvement over time can be seen. The figures for Design Group 1 show the false match rates for the first matching undertaken with live Census data. The figures for Design Group 2 relate to the first matching run after training and system changes. The figures for Design Group 3 show the false match rates for the first matching run after the implementation of the search protocol (see 5.6).

5. Changes made to the matching strategy

- 5.1 The false match rates shown in the previous section are all due to matcher error, to a greater or lesser degree. This section examines the recurring errors found in each of the matching runs shown in the table and the steps that have been made to avoid these errors in future.

Evaluation of Design Group 1

- 5.2 Differences between the addresses appearing in the CCS address table and the geography database were identified as a reason for 31% of the missed matches. Another 31% of false negatives were cross postcode matches. This occurs where there is a difference between the postcodes on the CCS and Census records. In 15% of false negative matches different names had been entered onto the CCS form and the Census form. In 10% of missed matches, scanning error had created a difference between the address or name details in the CCS and Census. 13% of missed matches appear to have no explanation and reflect human error only.
- 5.3 In order to remedy the problems identified above several changes were made to the matching system. During the first matching run of Design Group 1, supervisors did not have a specific quality assurance role. They tested a sample from each postcode to check for false positive and false negative matches. This was a time consuming process as the system was not really designed to facilitate this step. The most significant change to the matching strategy brought about by the analysis of the Design Group 1 results is the quality assurance role of the supervisors. The matching system now presents the matching supervisors with all unmatched records. This enables them to run Design Group wide searches and ensure that false negatives are kept to a minimum.
- 5.4 Additional training was also provided to the matching team to reduce human error. Matchers were re-trained in search techniques in order to assist them in finding potential matches. Different scenarios were also covered involving what to do when presented with poorly scanned forms. These steps reduced the false negative match rate on the initial match of an area from 1.32% to 0.28%, still short of the target of 0.1% overall.

Evaluation of Design Group 2

- 5.5 All the missed matches resulted from inadequate searching. Different individuals within the matching team had developed their own search techniques and some were more successful than others.
- 5.6 To solve the problem of the inadequate searches, a search protocol has been introduced which all members of the matching team must follow. This protocol is based on "best practice" taken from the searching procedures of the most successful matchers. This was accompanied by refresher training in searching and avoiding false positive matches. The searching protocol and training reduced the false negative match rate to 0.06% for the initial matching run of Design Group 3. This is below the target error rate of 0.1%.

Evaluation of Design Group 3

- 5.7 The three false negative matches found in the output from Design Group 3 are all a result of simple human error. Human error is the most difficult problem to eradicate. It would take only six false negative person matches, or around three households, to push the error rate above 0.1%.
- 5.8 The strategy adopted for reducing the scope for human error and therefore maintaining or improving matching accuracy is to try and reduce human intervention. The probability weights are currently being trained, based on the output of the first five Design Groups. The improved weights will enable an increase in automatic matching from the current rate of 45%. As the number of completed areas increases and the weights get more accurate, it is expected that the proportion of matches made automatically will increase to around 60 or 70%.

6. Conclusions

- 6.1 At present, all Design Groups are matched a number of times, allowing all identified false negative matches to be corrected. We can therefore be confident that all areas matched to date are well within the 0.1% false negative matching target set by Kendrick.
- 6.2 The changes to the matching strategy outlined in this paper have quickly reduced the false negative match rate to below the 0.1% target for each individual pass of the data. Therefore, should timetable pressures prevent us from double matching in future, we are confident of still achieving the necessary accuracy targets.