Development of pseudonymised matching methods for linking multiple administrative datasets

Pete Jones, Office for National Statistics

Keywords: record linkage, pseudonymisation, administrative data, census, population

1. Introduction

This paper discusses research to develop methods to link record level administrative datasets that have been pseudonymised at source. Matching multiple administrative sources is both resource intensive and elevates risks relating to the privacy of data about people and households. This paper outlines new approaches that have been developed to accurately link datasets that have been pseudonymised with a secure hashing algorithm.

It is well known that pseudonymising data prior to linkage inhibits the use of probabilistic matching algorithms and clerical resolution. Critical to the research has been the development of techniques that can identify similarity between pseudonymised matching fields and the use of statistical techniques to accurately designate candidate pairs into matches and non-matches. A quality assurance exercise has been undertaken to test these methods in a comparison of match results between the NHS Patient Register and the 2011 Census. Results so far are highly promising showing relatively low levels of false positives (less than 1%) and false negatives (around 2%). The paper will also discuss further ideas for improving the quality still further.

2. Methods

2.1 Deterministic Matching

The matching algorithm comprises three stages. The first is deterministic linkage based on a sequence of link-keys that are derived on each dataset prior to importing data into the linkage environment. Link-keys are based on concatenations of selected components of identifiers (i.e. names, dates of birth and addresses), resulting in unique identifiers for the majority of records on any dataset. These identifiers are then pseudonymised with the hashing algorithm and used to match records between datasets in the linkage environment.

The algorithm runs a sequence of pseudonymised link-keys, each of which attempts to resolve particular types of inconsistency that occur across match fields. Records linked uniquely are designated as matches, with the remaining residuals moved to the second stage of the algorithm.

2.2 Similarity Tables

Link keys are successful in identifying true matches between records that have relatively low levels of disagreement between match fields. For more complex cases involving disagreement across multiple match fields, measures of similarity between match fields are
required. A method for achieving this has been developed through the construction of similarity tables which are also constructed during pre-processing and before pseudonymisation.

Two de-duplicated lists of forenames and surnames occurring across the two datasets are compiled prior to pseudonymisation process. String comparison algorithms are undertaken between all names in the list to develop a thesaurus of names that are similar, with a comparison metric indicating the level of similarity. These tables are then pseudonymised and imported into the linkage environment and used to identify candidate match pairs between the two datasets.

2.3 Score based matching

From the records identified as possible matches from the similarity tables, a small sample of candidate pairs are made available for clerical matching (approx. 1000 pairs). A decision is made whether to designate each pair as a match or non-match. This clerically matched dataset now serves as training data to classify the match status of all other candidate pairs identified by the similarity tables. The current method relies on a logistic regression model, where the dependent variable is the binary outcome (yes or no) regarding match status. The predictors include the following: name similarity scores, date of birth similarity score, name commonality, sex agreement, postcode agreement and geographic distances between locations. Candidate pairs with a match likelihood >=0.5 are designated as matches, those below as non-matches.

2.4 Comparison Study

The pseudonymised matching algorithm has been tested and compared with the results of a previously undertaken linkage exercise, where links were identified using a combination of exact matching, probabilistic automatching, clerical resolution and clerical searching. In both cases a sample of NHS Patient Register records have been linked to the 2011 Census of England and Wales. The results of this matching exercise have been used as a measure of quality for the pseudonymisation algorithm on the basis that it is close to a ‘gold standard’ in identifying all of the true matches between the two datasets.

The pseudonymisation process and the methods developed are described in detail in ONS(2013).

3 Results

Table 1 shows the results of the comparison exercise. The match rates achieved by the pseudonymisation method are slightly lower than those achieved by the Census matching team. Eight local authorities were used as the basis for comparison, five of which were in cities (London and Birmingham), and three were in more rural areas. Matching scenarios are generally more complex in city areas owing to high rates of population migration and greater ethnic diversity in naming conventions. The table shows that the false positive rate is only slightly higher in city areas than more rural areas, however the increase in false negatives is more evident.
Table 1 – Match rate comparison and errors for pseudonymised matching method

<table>
<thead>
<tr>
<th>Local Authority Type</th>
<th>Census match rate</th>
<th>Pseudonymised match rate</th>
<th>Pseudonymised false positive rate</th>
<th>Pseudonymised false negative rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>City Local Authorities</td>
<td>71.3%</td>
<td>70.3%</td>
<td>0.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Rural Local Authorities</td>
<td>91.2%</td>
<td>90.7%</td>
<td>0.3%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Total</td>
<td>72.7%</td>
<td>71.7%</td>
<td>0.5%</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

4 Conclusions

The results of this comparison study indicate potential for a quality pseudonymised linkage approach. The method described has been implemented to produce trial outputs on the size and characteristics of the population from 2015 onwards. It may not be possible that pseudonymisation methods will meet the very high quality levels achieved through more conventional approaches. However research continues in this area, with a provisional target of lowering the false negative rate to less than 1 per cent.

The motivation for researching a pseudonymisation approach is critical to the development of future Census taking in the UK. Exploring administrative data as the basis for improving the quality of census outputs will be a major focus in preparations for the 2021 Census and research continues to investigate the potential of an administrative data based census to replace the traditional ten yearly census in England and Wales after 2021. In undertaking this research, linkages between administrative datasets will be made on a large scale (most datasets have tens of millions of records) and consideration for the associated risks to privacy need to be taken into account. This necessitates a different and more automated approach to be taken. There may be other solutions to alleviate the privacy risks in due course, but the scale of the matching in this context prohibits anything other than a highly automated approach.

References