Automatic Record Linkage of Individuals and Households in Historical Census Data

Zhichun Fu¹, Mac Boot², Peter Christen¹, Jun Zhou³

¹ Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
Canberra ACT 0200, Australia
{sally.fu, peter.christen}@anu.edu.au

² Australian Demographic and Social Research Institute
College of Arts and Social Sciences
The Australian National University
Canberra ACT 0200, Australia
mac.boot@anu.edu.au

³ School of Information and Communication Technology
Griffith University
Nathan QLD 4111, Australia
Jun.zhou@griffith.edu.au

1. Introduction

Historical census data captures information about our ancestors. They help social scientists to understand how our ancestors lived, as well as the economic, social, and demographic features of their society.¹ Invaluable as they are, census returns are still only snapshots of moments in time. The value of these snapshots is greatly enhanced, however, if they can be linked to the same individuals, families, and households over several censuses. Linked censuses can provide social scientists with new insights into the dynamic character of social, economic, and demographic change, and enable researchers to reconstruct the key life course events of large numbers of individuals, households, and families at levels of detail far beyond the scope

of traditional methods of historical research.^{2:3} They may even facilitate epidemiological studies of the genetic factors of diseases such as cancer, diabetes, or mental illnesses.⁴

Social scientists have linked census records manually, but the process is very expensive in terms of time and human resources, and such exercises are usually restricted to small numbers of individuals and households over short periods of time.^{2;3;4;5;6;7;8} The most promising way of tackling these problems is through the development of automatic record linkage techniques. The main problem with these techniques is how to eliminate ambiguous links that appear to be true (i.e. two records that refer to the same individual) but turn out to be false. Ambiguous links arise for three main reasons: the poor quality of the census records themselves; the populations that the censuses attempt to record are themselves highly dynamic (people are born, they marry and form new families, their addresses and occupations change, they migrate, and they die); and because many people have the same or very similar names. While data cleaning and enhancement may cure data quality problems, population dynamics and common name problems are not so easily solved, and the difficulty of eliminating ambiguous links still remains as the chief obstacle to achieving highly accurate record linking results.

This paper introduces a novel method that unifies individual and household linkage which we believe will help to improve the accuracy of record linking in historical census data collections. The process involves three steps: First we clean, standardise and enhance our census data. Second, we use either a supervised machine learning technique (which learns a model based on a set of true matching and non-matching record pairs), or a similarity threshold method (where the match decision depends on whether two records have a similarity that is above a preset threshold) to link individual record pairs across

census datasets. Third, we use the identified matched pairs to perform household linking that considers the matching of several, or all, members of each household rather than treating each person as a single isolated individual. This step allows us to take advantage of a rich field of information that effectively removes many ambiguous links and enables us to perform more accurate record linkage than is possible by using traditional methods. It is the final step that provides the novel feature of our paper.

2. Previous Attempts to Link Historical Census Data

Social scientists have linked census records for decades using both manual and automatic methods. Larsen, for example, looked at the problem from a probabilistic point of view using a maximum likelihood estimation model to separate record pairs into possible matches and non-matches.⁹ Manual checking was then performed to update the estimation model. This process was iterated until few additional matches remained. Ruggles attempted to limit false matches by selecting attributes that did not change over time. Ambiguous links were then removed to achieve high rates of linkage accuracy.³ The effect of this method was to scarify a large number of links including many correct links. Vick et al. standardised name strings in a population study of census data from the United States and Norway.^{10:11:12} The authors used name dictionaries and estimates of name frequencies to select how name values are to be cleaned and standardised. The Jaro-Winkler approximate string comparison algorithm was then used to match candidate names to their standard form.¹³ The effectiveness of the standardisation was validated by the fact that it greatly reduced the number of false links.

The attributes used in these methods vary greatly depending on what detailed personal information is available in the census dataset(s) under study. The

most commonly used attributes include first and last names, house number, street name, phone number, age, birth year, birth place, parent's birth place, relation to head of household, marital status, sex, and race. The sizes of the datasets used in these approaches also vary a lot. For example, Larsen and Rubin validated their method on five US Census/Post-Enumeration Survey Datasets with more than 288,414 links containing 26,315 matches.⁹ Ruggles, on the other hand, performed research on 500,000 individuals in the Church of Jesus Christ of Latter-Day Saints database, and one per cent of the Integrated Public Use Microdata Series.³

Existing efforts to improve linkage accuracy are mostly focused on standardising names, removing ambiguous links, or combining different attributes to improve similarity scores for the compared record pairs and to deal with low data quality problems (discussed below) that have seriously hampered the improvement of linking accuracy in previous works. For example, name standardisation methods developed by Vick and Huynh have increased the number of single matches but have failed to reduce the problem of multiple ambiguous links.¹⁰ The result is that many methods simply discard ambiguous links, which leads to the loss of large amount of potentially true matches, or have left these for manual checking.^{3:9:14}

Goeken et al. attempted to deal with the inaccuracy of 19th century census data by generating initial linkage results using name and age similarity scores, name commonality, and birthplace distribution measures.¹⁴ They then use the single record links that have a very high confidence value as primary links to identify matched households. Once matched, the linked households then allow them to assume that other members resident in both households are also confirmed positive matches even though they might have low similarities that would otherwise have been treated as ambiguous and therefore be

rejected. The distinct feature of their approach, however, is that even though they use household information, their linking step remains dependent on the initial single record link. Our approach, which is introduced in Section 4, deals with the difficulties generated by the poor census data by treating the household as an integrated unit. We use ‘whole of household’ similarities (a very rich data source) to identify and confirm any individual record link that is not a false positive (i.e. is a true link) and similarly to overcome the ambiguity problem, by using whole of household similarities to check that other members resident in both linked households are also confirmed positive links. In this way we retain all record cases as potential true links thus providing a broader base upon which to build more complete record of linked person records over several censuses.

3. The Lancashire Censuses of 1841 to 1901

We use six census datasets covering the district of Rawtenstall in North-East Lancashire between 1851 and 1901. The original hand-written returns (a sample image is shown in Figure 1) have been scanned into digital form and manually transcribed into tables and stored on spreadsheets. The spreadsheets contain records of the name, address, age, sex, marital status, relationship to the head of household, and occupation, along with other data for each person in the district. All spreadsheets contain at least 12 attributes for each record (see Table 2). In total, the six datasets contain around 160,000 records. Table 1 lists the numbers of records in each dataset.

3.1 Data Quality Problems

The poor quality of the original census returns and of some modern transcriptions is notorious. The census enumerators who delivered the householder schedules to the household were responsible for ensuring that the returns were filled in correctly. The schedules contained detailed instructions on how they should be filled in these details, but low levels of

literacy left many households confused and unable to complete the returns correctly.

The enumerators were instructed to check the schedules on the doorstep when they collected them. Unfortunately many were themselves unclear in their own understanding of the instructions, or were slipshod in their checking, or misunderstood householder replies. The enumerator then transcribed the data from the schedules onto enumerator returns sheets. At this point corrections were made if necessary, though the process also provided opportunities for introducing new errors. More errors were added as the enumerator returns were passed through various supervisors to a central office where the data were extracted for use in the census reports. Extraction was completed by clerks temporarily appointed for the task who were paid by the number of data items they abstracted, as measured by score marks they struck through the items in the enumerator returns. They were also required to add various corrections into the enumerator's return sheets. These marks and corrections were in various coloured inks and pencils that differentiated them from the written text. Later, when the census images were digitised, colours were lost causing the marks to obscure many data items.

Together, the poor handwriting, the errors introduced during the enumeration and consolidation stages, and obscured data caused much further compounding of the errors when the data contained on the census images came to be transcribed onto computer spreadsheets. An additional problem is that the process of transcription, even for a small community like Rawtenstall, involves hundreds of thousands of key strokes that inevitably introduce numerous opportunities for errors. In census data where some items consist of a single keystroke, error risk becomes a serious problem.

Given these data quality problems, the complications of common names, the presence of limited and non-standard information in key attributes (fields), and the opportunities for population dynamics to transform populations significantly during the ten year intervals between censuses, it is not surprising to find that attempts to use traditional automatic record linking techniques to deal with large collections of census data have been disappointing.

4. Linking Model

In this section we describe our approach to automatically linking historical census data. We start by providing an overview of our approach, which is shown schematically in Figure 2, we then describe each step in detail.

4.1 Overview

We consider the household as the basic reference group for linking every pair of individuals in consecutive census datasets. Inputs to the system consist of two or more census datasets to be linked; the outputs consist of individual record pairs and household pairs that have been classified as matches between two adjacent censuses across all the datasets.

The approach comprises six steps as illustrated in Figure 2. The first step is an interactive manual data exploration process that provides direct hints for the cleaning step. The second step focuses on data cleaning and standardisation to resolve the problem of low quality data in the census data. Its purpose is to enhance the data quality by identifying missing, wrong, and inconsistent values, and by transforming the data to a standardised form, thereby increasing the likelihood of finding true matches. Step 3 identifies each household and assigns it a unique household identifier (HID). Among other uses these HIDs are employed to clean the address attributes. Step 4 uses a blocking technique to subdivide a dataset into blocks to limit the

number of record pairs that will be compared with each other to those records in the same block. This action speeds up the linkage process significantly.¹⁵ Using the appropriate blocking criteria ensures the linkage rate is affected only to a minor degree. The fifth step is pairwise identification of possible matched record pairs across the census datasets. This involves record link classification processes that are performed using either a similarity threshold method, or a supervised machine learning method. The final step is the group linking step that reduces the ambiguity of both individual record linkage and household linkage results.

The development of each step is guided by domain knowledge, particularly, household information on which basis the original census data were collected. Detecting all households and attaching a householder identifier is an essential step toward successful census record linkage because household identifiers provide a wealth of information. This information is invaluable in the linkage step, where it improves the number of confirmed linked record/household pairs significantly. More detailed discussions about our approach are presented in the following sections.

4.2 Data Exploration

Before entering on to data cleaning and standardisation, we used the Febrl record linkage tool to identify missing, inconsistent, and wrong values.¹¹ The results of this exercise for the 1851 census data are summarised in Table 2. As can be seen cleaning reveals that there were large numbers of missing values as well as anomalous values in the census dataset. For example, the number of unique values in the SEX attribute should be two — female ('f') and male ('m') — not three. Further analysis shows the third value as 'j', which is not meaningful and is removed in the data cleaning step. Another example is the values in the AGE attribute, which is a mixture of digits and letters

implying that the original values were entered in different formats and therefore data standardisation is required. Several attributes contain errors and values in mixed formats as, for example, the FIRST NAME attribute which contains digits, letters, punctuation, and other symbols. Such attributes require cleaning and standardisation to be applied. Note that the numbers of data entries with missing value in some attributes have increased after the cleaning step. This is because some non-meaningful values such as “not identified” have been removed.

4.3 Data Cleaning and Standardisation

Data cleaning and standardisation operations have to be applied to improve quality and formatting of the data to a consistent form. To do this, we use the method proposed by Christen, which requires the following operations.¹¹

- All values are converted into lower case letters.
- First and middle names are split into two attributes.
- Age standardisation is applied to convert the age format into a digits-only format, which represents an age as a number of years, using a purposely built lookup table to convert special values, such as converting ‘under 1 month’ into a year value of ‘0’.

To correct inconsistent values in a single record, look-up tables were used that contain correct value pairs. For example, when processing values in the SEX attribute, each entry is validated by the RELATIONSHIP TO HEAD OF HOUSEHOLD attribute using a matching table defined to map entries to SEX options such as from ‘daughter’ to ‘female’. Thus in our datasets, the unique values of SEX changed from three to two after cleaning, and AGE values

changed to numbers ranging from 0 to 92. To prevent loss of information during cleaning, some values, such as '1/2 scholar and 1/2 piecer', have to remain unchanged. Table 2 shows the changes of attributes before and after the data cleaning and standardisation step.

4.4 Automatic Household Identification

We adopt the following assumptions when automatically identifying households: 1. that each census form was filled in by the head of household, and 2. that the relationship to the head of household attribute for a household always starts with the head of household. This allows us to build a set of rules to apply when identifying households and the relationship of its members to the head. Four possible values define the head of a household's relationship: 'head', 'head of family', 'widow', and 'widower' (see Table 3). Using this table we developed a search algorithm to scan through a census data file. Each time a record has one of the head of household values the Household Identifier (HID) is incremented by one and is assigned to all following records until another record with a head of household role is found. An example of such searching result is shown in Table 3. The detected number of households in each dataset is given in Table 1.

The HID detection results can be used for further cleaning of the census data. Here we show how they can be used to reduce discrepancies in addresses of members in a household. To do this we use the HID to separate the relationship of home address and surname in every household into four categories: 1. same address with same surname; 2. same address with different surnames; 3. different addresses with same surname; and 4. different addresses with different surnames. Figure 3 shows about 70 per cent of households in each census fall into the first category, which suggests correct cleaning. About 30 per cent of households contain records that have

the same address but different surnames, which can be explained by the fact that population growth in Rawtenstall continued to outstrip the growth in housing space throughout our period requiring large numbers of short term residents to board with other families, and couples in the early years of marriage to live with parents, rather than errors made during the data cleaning stage.

We are mostly interested in identifying households with different addresses but same surname, which may be caused by errors in the data entry process where operators have entered data in a wrong row from the original table. For these cases, a program was developed to find the address that was common to most household members. Different addresses are then automatically converted to the common address. To avoid introducing errors we make such conversions only when there are no more than two different addresses that appear in one household, and the majority of household members have the same address.

4.5 Linking Attributes

Some of the available 12 attributes cannot be used in record linkage. The IMAGE REF attribute, for example, indicates the location of the record in the scanned image database. We use the following six attributes in the record comparison step: SURNAME, FIRST_NAME, SEX, AGE, ADDRESS, and OCCUPATION. Among these, FIRST_NAME and SEX are less likely to change across time though FIRST_NAME may be reported differently when, for example, a diminutive such as 'Liz', 'Betty', or 'Beth' is used instead of 'Elizabeth'. AGE should normally accrue by 10 years (and never less than 9 or more than 11) between two consecutive censuses. SURNAME will change when a female marries, while ADDRESS and OCCUPATION may or may not change.

A major challenge is the dynamic character of household and individual record changes between censuses: children are born, people die, adults marry, occupations change, people retire, change address, or migrate; household membership may also change, perhaps unrelated to the family, as boarders, guests, servants, or lodgers enter or leave. Addresses also become increasingly non-specific as we go back into the nineteenth century. This variability makes the reconstruction of family and household data across time even more difficult.

4.6 Pairwise Linking

We implemented the pairwise linking process using both a support vector machine (SVM) method and a similarity threshold method.^{17,18} SVM is a supervised learning technique that learns a mathematical model from a labelled training dataset and applies this model to all unseen data. All data contained in our application are in the form of record pairs that are classified as being a true match (referring to the same individual) or a non-match (referring to two different individuals).¹⁶ The similarity threshold method make a decision by separating data based on how similar they are.

Supervised learning usually achieves higher accuracy results because it learns a model using manually annotated data (ideally provided by a domain expert) which provide ground truth information about the data. The similarity threshold method is more commonly used because manually provided labels (match/non-match) are often difficult to collect on large datasets.

We apply a blocking technique to reduce the complexity of pairwise linking. The linking process is very time consuming when large datasets are used because potentially all pairs of records from both datasets have to be compared. Blocking aims to solve this problem by dividing datasets into smaller blocks that limit the number of potential record links to those in the same block.¹⁵ Blocks are generated using criteria commonly known as 'blocking keys'. These consist either of a single attribute or a combination of attributes. Attributes with good data quality and uniform data distribution were selected to form blocking key(s). Only records with the same value in their blocking key are compared.

We apply the following three blocking keys (BK):

- SURNAME concatenated with SEX and CENSUS_PARISH (BK1)
- FIRST_NAME concatenated with ADDRESS (BK2)
- INDUSTRY CODE concatenated with SURNAME (BK3)

When comparing records, appropriate approximate string comparison functions were chosen for each attribute.¹⁸ The list of attributes and functions used to compute the similarities between values is shown in Table 4. Details on these functions can be found in Christen, 2007.¹⁹ For each compared attribute, a similarity score of between 0 and 1 is generated. The higher the score the more similar are the two attributes (scores of 1 indicate an exact match; 0 means no similarity). For any given record pair, the sum of the attribute-wise similarity scores reflects the overall similarity between the record pair, with the similarity increasing monotonically in accordance with the summed value. This allows the adoption of a similarity threshold method to classify pairs of records into matches (if their sum is above the selected threshold value) and non-matches (if their sum is below the threshold).

Appropriate setting of the threshold influences the accuracy of the final matching results. To find that setting we compare the linked results with respect to different threshold values (ρ). We found that the number of 'single links' (i.e. a record in one dataset that is linked to only one record in another dataset) and the number of 'multiple links' (i.e. a record in one dataset that is linked to several records in another dataset) intercept at certain points. Figure 4 shows how the threshold trends in the number of single linked records intercept with the number of multiple linked records. The single linked record thresholds are represented by solid curves and the multiple linked thresholds are indicated by dashed lines. We use these intercept points as default threshold values for linking census datasets. Therefore, the threshold is set to 4.0 for the linking of all pairs of our census datasets because the interception points in all the six pairs of datasets are remarkably consistent. Each intercept means that there are more single link pairs with similarity values above the threshold compared to record pairs with multiple links. We consider as matched only those record pairs whose total similarities are higher than the threshold. Table 7 shows the total number of matched record pairs between each pair of census datasets, as well as the number of distinctive records in the first dataset with single link and multiple links.

We also used the SVM method for pairwise linkage. SVM takes a set of input data as training set to train a model, and then applies the model to new record pairs to predict which of them are matches or non-matches. The input to the SVM model contains class labels and attributes vectors whose entries are the attribute-wise similarity scores calculated using the functions summarised in Table 4. To generate the class labels for data, we manually labelled 4,000 matched record pairs that we are highly confident of to correspond to the same person in the 1871-1881 datasets. By using the labelled data we can

automatically generated negative training samples, which are non-matched record pairs.

We performed two experiments to validate the effectiveness of the pairwise linking methods. The first experiment generated a dataset with 4,000 positive samples and 46,000 negative samples. These data were randomly split into a training set and a testing set in three settings, 10:90, 50:50, and 90:10, respectively. The parameters of the SVM were tuned to be optimal. The default setting for the similarity threshold method is $\rho=4.0$ based on a previous data analysis.²⁰ For each train/test split setting, we ran the experiments ten times and report the means and standard deviations of accuracy, precision (number of true positives divided by the total number of positive results), and recall (number of true positives divided by the total number of records that actually belong to the positive class).¹⁸ The results are summarised in Table 5. These results show that our methods performed exceptionally well, with more than 97 per cent classification accuracy for both SVM and the similarity threshold method. It also shows that the SVM method achieved higher than 95 per cent precision and recall, which significantly outperformed those of the threshold method. The SVM method has, therefore, a clear advantage over the similarity threshold method on a small dataset. It is interesting to see that in different data split settings, the overall performances of both methods do not change much. This suggests that we can use a small number of positive samples to train a final pairwise record link classification model.

To demonstrate the performance of the proposed methods on larger amounts of data, we generated a dataset with 400 positive samples and 314,445 negative samples in the second experiment. This data distribution is closer to the true data distribution - it has a small number of matched record pairs but a

very large number of non-matched pairs. The classifiers were trained and tested on a 50:50 data splitting, with results given in Table 6. Similar to the first experiment, both methods have achieved very high classification accuracy. The SVM method has significantly outperformed the threshold method in recall, with 97.98 ± 0.47 per cent and 85.25 ± 1.59 per cent, respectively. This suggests that most true links can be found. On the other hand, the precision of both SVM and the threshold method are lower than 10 per cent, which suggests that many non-matches have been classified as matches, i.e., the results contain large numbers of multiple links that are ambiguous. Further data processing is therefore required to reduce these ambiguous links. This can be achieved by the following group linking step. Finally, in Table 8, we report the results of the total number of matched record pairs between each pair of census datasets using the same SVM model trained above and the default $p=4.0$ for the similarity threshold method.

4.7 Group Linking

Machine generated pairwise linking results have been widely used by social scientists as the final outcomes of record linkage exercises, and they have used the results for further investigations. As shown from the previous experiments, however, the results contain large numbers of multiple links that need further investigation before they can be accepted as true matches of a single record to a record in another census return. This usually requires much manual effort to identify which linked record pairs are true matches, effort that is time consuming, cumbersome and error prone.

We argue that the danger of generating numerous multiple links can be solved if the relationship between household members is taken into consideration. A simple example is shown in Figure 5 where 'Sarah TAYLOR', in the middle household on the left-hand side, is matched to two records with

a similar name in two different households on the right-hand side. In this example the middle family on the right-hand side is obviously the true match. Another example is 'Daniel TAYLOR' in the lower right hand panel, who is linked to two different persons on the left hand side with the true match found in the left-hand lower panel. In both cases the true match is obvious because they are read in the context of other family members.

Based on this observation, we propose taking household information into account. This allows several linked records belonging to the same household to be grouped, which is then used to calculate the best unique pairs in households to match across the census datasets. In this way household linking utilises richer information than would be used in standard pairwise linking procedures, which leads to refined selection of the correct record links while generating correct household links simultaneously. In this way we reduce both linking ambiguity and increase linking accuracy significantly.

Group linking was originally developed for bibliographic data applications.²¹ An advantage of the method is that not all members in a group need to be matched, which exactly meets the requirements of household matching where families may have changed significantly between two censuses, and where some members cannot be matched to anyone in the previous or next census. For this purpose we modified the original group linking method to be suitable for linking record pairs or household pairs in a way that generates group similarity scores for each pair of households. These scores can be calculated using the sum of normalised similarity of the matched individual record pairs in two households. We did not consider children with ages less than nine years in the second of two census datasets because they had not been born when the first census was collected. This allows us to be certain that matches of these records cannot be found. Finally, when a record in the first dataset is

matched to records from multiple households in the second dataset after the pairwise linking step, the decision on which record should be the true match can be made by selecting the record in a household of the second dataset that has the highest household similarity.

In summary, after completing the pairwise linking step we perform household linking for each record in the first census dataset that is linked to more than one record in the second census dataset as follows:²²

- 1) calculate the number of members in each household;
- 2) calculate the sum of the similarity score for the linked record pairs between the household pairs;
- 3) compute the household similarity by dividing the resulting sum of similarity scores by the number of distinct members in both households;
- 4) Identify the household pair and corresponding record pairs with the highest household similarity.

The group linking results are shown in Table 7 and 8. It can be seen that the number of records with multiple links has been greatly reduced. This implies that group linking is very effective. Table 9 shows the number of households with single and multiple links after group linking has been applied to records with multiple links. There will still be records with multiple links even after the group linking is completed. This is because group matching of several households inevitably generates close group similarity scores. Furthermore, because families and households can change dramatically between censuses, low household match scores do not necessarily mean that two households are not matched. In these cases it is hard to tell which household is the true match unless further analysis is conducted on the household and families concerned. At this stage, we assume all household links found are the true

matches, and have not set thresholds to eliminate possible false matches as was done in the original group linkage approach.²¹

5. Discussion

The core idea introduced in this paper is that linking both individual and household records using a group linkage method greatly improves the elimination of ambiguous links and achieves highly accurate true linkage results. Group linking aims at removing as many multiple links as possible and does not reduce the number of records with single links. The method takes domain knowledge of household information into account to improve the record linking results. Household similarity scores help a researcher to judge which household and record pair(s) are most likely to be correct links. The results show that the method reduces the number of ambiguous links very significantly and increases the accuracy of the linkages. In these ways it provides the social scientist with a feasible tool to analyse historical census data on a larger scale than has hitherto been possible using alternative record linking procedures.

It is useful to see the value of these tools in the context of an historian or social scientist faced with the problem of linking individuals or households over several censuses: it is intimidating, even when attempting no more than linking them from one census to the next. To attempt the same task over six censuses for a whole town population is truly forbidding and any tool that automates the task is truly welcome. Using the proposed record linking method to link persons and households living in Rawtenstall, Lancashire over five census periods between 1851 and 1901 gave us single matches of between 23 per cent and 28 per cent and multiple matches of between 30 per cent and 47 per cent of all the persons living in the town. What is encouraging is that by adding the household identification and linking procedures

described, the single matches are increased to a range between 44 per cent and 56 per cent, while the multiple matches are reduced to the range between 3 per cent and 8 per cent. It also shows that household matches are possible for up to 82 to 90 per cent of all households in the community. Such large numbers over such a long period can only enrich the prospect of gaining new and richer insights into the lives of the people who lived in those days and the events and processes that affected them.

6. Conclusions

In this paper, we have introduced a group record linkage method for automatic individual and household linking in historical census data. The goal of this research is to reduce ambiguous matches by eliminating multiple record and household links. The key idea is that members in a household are considered as a group. After similarities between record pairs are computed, they can be classified as matches or non-matches using either an SVM classifier or similarity threshold method. These similarities and record link classification results can then be used to generate household linking similarities. Multiple record links that are presented in households with low similarities are then removed. We have tested our method on six Rawtenstall datasets. The results show that the proposed method effectively reduces the number of multiple record and household matches and provide social scientists with a useful tool to process and analyse historical census data.

We plan to explore the relationship between individuals in a household, and to use graph methods to match structures of households. We will also apply the proposed methods to other dataset, which can be done by following the steps listed in Figure 2.

Endnotes

1. E. Higgs, *A clearer sense of the census: Victorian censuses and historical research*, (London, 1996).
2. D. Quass and P. Starkey, 'Record linkage for genealogical databases', *Proceedings of the 9th ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, (Washington DC., 2003), 40–42.
3. S. Ruggles, 'Linking historical censuses: a new approach', *History and Computing*, 14(2006), 213–224.
4. E. Glasson, N. de Klerk, A. Bass, D. Rosman, L. Palmer, and C. Holman, 'Cohort profile: The western Australian family connections genealogical project', *International Journal of Epidemiology*, 37(2008), 30–35.
5. G. Bloothoof, 'Assessment of systems for nominal retrieval and historical record linkage', *Computers and the Humanities*, 32(1998), 39-56.
6. G. Bloothoof, 'Multi-source family reconstruction', *History and Computing*, 7(1995), 90–103.
7. E. Fure, 'Interactive record linkage: The cumulative construction of life courses', *Demographic Research*, 3(2000), 11.
8. A. Reid, R. Davies and E. Garrett, 'Nineteenth century Scottish demography from linked censuses and civil registers: a “sets of related individuals” approach', *History and Computing*, 14(2006), 61–86.
9. A. Larsen and D.B. Rubin, 'Iterative automated record linkage using mixture models', *American Statistical Association*, 79(2001), 32-41

10. R. Vick and L. Huynh, 'The effects of standardizing names for record linkage: Evidence from the United States and Norway', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(2011), 15–24.
11. P. Christen, 'Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System', *SIGKDD Exploration*, 11(1) (2009), 39-48.
12. Minnesota Population Center: <http://www.ipums.org/>
13. W.E. Winkler, Overview of research linkage and current research directions, US Bureau of the Census, Statistical Research Report, (2006).
14. R. Goeken, L. Huynh, T. A. Lynch, and R. Vick, 'New methods of census record linking', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44(2011), 7–14.
15. P. Christen, 'A survey of indexing techniques for scalable record linkage and deduplication', *IEEE Transactions on Knowledge and Data Engineering*, (2012). vol 24 (no 9), pages 1537-1555
16. C. Bishop, *Machine learning and pattern recognition*. (New York, 2006).
17. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed, (2006)
18. P. Christen, *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer (2012).
19. P. Christen, 'A two-step classification approach to unsupervised record linkage', *Proceedings of the 6th Australasian conference on Data mining and analytics*, (2007), 111-119.

- 20.Z. Fu, P. Christen, and M. Boot, 'Automatic cleaning and linking of historical census data using household information', *Proceedings of the 2011 Workshop on Domain Driven Data Mining*, (2011), 413-420.
- 21.B.W. On, N. Koudas, D. Lee, and D. Srivastava, 'Group linkage', *Proceedings of the IEEE 23rd International Conference on Data Engineering*, (2007), 496–505.
- 22.Z. Fu, J. Zhou, P. Christen, and M. Boot, 'Multiple Instance Learning for Group Record Linkage', *Proceedings of the Sixteenth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'12)*, (2012), 171-182.

Figures and Tables

Figure 1: A sample original image

Civil Parish (or Township) of	City or Municipal Borough of	Municipal Ward of	Parliamentary Borough of	Hamlet of					
Holy Trinity	Hungate upon Hull	South of the Hungate upon Hull	Hull	St. Barnabas					
No. of Schedule	ROAD, STREET, &c., and No. or NAME of HOUSE	HOUSES In-cluded (A), or Excluded (B)	NAME and Surname of each Person	RELATION to Head of Family	CON- DITION as to Marriage	AGE last Birthday of	Rank, Profession, or OCCUPATION	WHERE BORN	(1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
1136	1144, Pearson Lane	1	James Ward	Lodger	Married	38	Engine driver	Lincolnshire, Wigan	
			William Brown	Head	Married	39	Bookbinder unemployed	York, Hull	
			James E. Do	Wife	Married	38		York, Hull	
			Arthur E. Do	Son	Married	16	Bookbinder	York, Hull	
			Alice E. Do	Daughter	Single	13		Do	Do
			Eliza E. Do	Daughter	Single	8		Do	Do
			David W. Do	Son	Single	7		Do	Do
115	5 Do	1	William Collins	Head	Married	26	Fireman	Lincolnshire, Rotherham	
			James E. Do	Wife	Married	25	Domestic	York, Sheffield	
			Ernest E. Do	Son	Single	13		Do	Hull
			Charlotte E. Do	Daughter	Single	11	Schooler	Do	Do
			Edith E. Do	Daughter	Single	8		Do	Do

Figure 2: Historical census record linkage processes

Figure 3: Distribution of household name and address relationships

Figure 4: Record linking results using the similarity threshold method

Figure 5: Illustrative example of multiple links.

Table 1: Total number of records and households in Rawtenstall datasets.

	1851	1861	1871	1881	1891	1901
Number of records	17,033	22,429	26,229	29,051	30,087	31,059
Number of households	3,295	4,570	5,575	6,025	6,378	6,841

Table 2: Record attributes and data quality before and after cleaning 1851 Census data set

This table shows the data quality analysis before and after data cleaning. “Unique” is the total number of data with unique values in an attribute, “Missing” is the total number of data with empty values in each attribute.

Attribute	Before data cleaning			After data cleaning		
	Unique	Missing	Type	Unique	Missing	Type
IMAGE-REF	929	0	mixed	929	0	mixed
ADDRESS	520	318	mixed	518	878	mixed
CENSUS-PARISH	10	0	mixed	10	0	mixed
COUNTY	1	0	letters	1	0	letters
SURNAME	1,028	0	mixed	1,024	5	mixed
FIRST-NAME	1,007	0	mixed	651	11	mixed
REL-HSEHLD	60	0	mixed	55	23	letters
SEX	3	0	letters	2	0	letters
AGE	129	1	mixed	91	1	digits
OCCUPATION	2,768	22	mixed	2,766	4,842	mixed
BIRTH-PARISH	1,406	5	mixed	1,403	323	mixed
BIRTH-COUNTY	52	3	mixed	53	7	letters

**Table 3: Sample electronic dataset for Rawtenstall
after cleaning and standardisation**

Surname	First name	Relation to Head of Household	Sex	Age	Occupation	Household ID
hey	frank	head	m	31	throstle overlooker cotton mill	3867
hey	annie	wife	f	30	-	3867
hey	willie	son	m	7	-	3867
hey	frank	son	m	5	-	3867
hey	hilda	daughter	f	1	-	3867
hey	annie	daughter	f	5m	-	3867
hargreves	ellen	widow	f	65	-	3868
hargreves	daniel	son	m	30	coal miner	3868
hargreves	samuel	son	m	25	coal miner	3868
hargreves	letita	daughter	f	23	cotton card room hand	3868
hargreves	james henry	head of family	m	34	cotton spinner	3869
hargreves	emma l	wife	f	35	-	3869
hargreves	john	son	m	9	-	3869
hargreves	fred	son	m	6	-	3869
hargreves	nellie	daughter	f	3	-	3869
hargreves	sarah olive	daughter	f	1	-	3869

Table 4: Similarity computation methods for the six attributes under consideration

Attribute	Methods ¹⁸
SURNAME	Q-gram/Jaccard
FIRST NAME	Q-gram/Jaccard
SEX	String exact match
AGE	Absolute value differences
OCCUPATION CODE	Percentage value differences
ADDRESS	Q-gram/Jaccard

Table 5: Pairwise record linking results on labelled data

This table shows the pairwise record linking results from both SVM and similarity threshold method on three data split settings.

Split	SVM method (%)			Similarity threshold method (%)		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
10-90	99.41±0.03	96.72±0.32	95.94±0.36	97.98±0.02	88.73±0.18	85.67±0.17
50-50	99.41±0.04	96.58±0.55	96.03±0.33	97.92±0.08	88.35±0.67	85.32±0.43
90-10	99.42±0.09	96.55±1.01	96.18±0.67	97.94±0.12	88.30±1.10	85.65±1.83

Table 6: Pairwise record linking results on large dataset

This table shows the pairwise record linking results from both SVM and similarity threshold method on a dataset with 800 positive samples and 314,445 negative samples.

	SVM method (%)			Similarity threshold method (%)		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
50-50	98.33±0.37	7.27±1.94	97.98±0.47	98.98±0.03	9.80±0.35	85.25±1.59

Table 7: Record linking results on six historical census datasets using the similarity threshold method

	1851-1861	1861-1871	1871-1881	1881-1891	1891-1901
Before group linking (Similarity threshold method)					
Total matched record pairs	57,961	68,566	70,307	62,312	65,449
Records with single match	4,079 (24%)	5,232 (23%)	7,061 (27%)	7,857 (27%)	8,315 (28%)
Records with multiple match	6,738 (40%)	8,492 (38%)	9,166 (35%)	8,786 (30%)	9,299 (31%)
After group linking (Similarity threshold method)					
Records with single match	7,618 (45%)	9,885 (44%)	12,486 (48%)	13,294 (46%)	14,215 (47%)
Records with multiple match	1,417 (8%)	1,802 (8%)	1,585 (6%)	1,299 (4%)	1,478 (5%)
Reduction of Multiple matches	79%	79%	83%	85%	84%

Table 8: Record linking results on six historical census datasets using the SVM classification

	1851-1861	1861-1871	1871-1881	1881-1891	1891-1901
Before group linking (SVM method)					
Total matched record pairs	56,301	71,752	80,802	80,504	79,442
Records with single match	4,163 (24%)	5,586 (25%)	7,053 (27%)	8,165 (28%)	8,465 (28%)
Records with multiple matches	8,008 (47%)	9,893 (44%)	11,489 (44%)	11,721 (40%)	11,871 (39%)
After group linking (SVM method)					
Records with single match	9,096 (53%)	12,058 (54%)	14,722 (56%)	15,956 (55%)	16,600 (55%)
Records with multiple matches	517 (3%)	649 (3%)	793 (3%)	876 (3%)	814 (3%)
Reduction of Multiple matches	94%	93%	93%	93%	93%

Table 9: Number of households identified with single (S) and multiple (M) matches for all linked datasets after group linkage, and percentage among all households in the first year.

	1851-1861		1861-1871		1871-1881		1881-1891		1891-1901	
	S	M	S	M	S	M	S	M	S	M
Similarity threshold method	2,889 (88%)	25 (1%)	3,896 (85%)	37 (1%)	4,671 (84%)	39 (1%)	4,951 (82%)	22 (0%)	5,275 (83%)	12 (0%)
SVM Method	2,956 (90%)	5 (0%)	4,035 (88%)	7 (0%)	4,886 (88%)	9 (0%)	5,208 (86%)	2 (0%)	5,478 (86%)	3 (0%)