A Supervised Learning and Group Linking Method for Historical Census Household Linkage

Zhichun Fu[1] **Peter Christen**[1] **Mac Boot**[2]

[1] Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
Canberra ACT 0200, Australia

[2] Australian Demographic and Social Research Institute
College of Arts and Social Sciences
The Australian National University
Canberra ACT 0200, Australia

Email: {sally.fu, peter.christen, mac.boot}@anu.edu.au

Abstract

Historical census data provide a snapshot of the era when our ancestors lived. Such data contain valuable information that allows the reconstruction of households and the tracking of family changes across time, allows the analysis of family diseases, and facilitates a variety of social science research. One particular topic of interest in historical census data analysis are households and linking them across time. This enables tracking of the majority of members in a household over a certain period of time, which facilitates the extraction of information that is hidden in the data, such as fertility, occupations, changes in family structures, immigration and movements, and so on. Such information normally cannot be easily acquired by only linking records that correspond to individuals. In this paper, we propose a novel method to link households in historical census data. Our method first computes the attribute-wise similarity of individual record pairs. A support vector machine classifier is then trained on limited data and used to classify these individual record pairs into matches and non-matches. In a second step, a group linking approach is employed to link households based on the matched individual record pairs. Experimental results on real census data from the United Kingdom from 1851 to 1901 show that the proposed method can greatly reduce the number of multiple household matches compared with a traditional linkage of individual record pairs only.

Keywords: Historical census data, household linkage, support vector machine, classification, group linking.

1 Introduction

Historical census data contain valuable information on individual persons and households at a given point in time. Such data allows us to reconstruct key aspects of households and families, such as birth, age, marital status, death, occupation, neighbourhood, and so on, that are of enormous value to genealogists, historians, and a wide range of other social and health scientists (Quass & Starkey 2003, Ruggles 2006, Glasson et al. 2008). As valuable as they are, these data provide only snapshots of the main characteristics of the stock of a population, capturing a vague image of how that stock and its characteristic features changed over time. To capture these changes requires that we link person by person and household by household from one census to the next over a series of censuses, a problem that hitherto has proved prohibitively expensive in time and human resources even for small groups of households (Anderson 1971). Once linked together, however, the census data are greatly enhanced in value. The linked results allow us to trace the changes in the characteristics of individual households, families and individuals over time. Linked information facilitates improved retrieval of information, and provides new opportunities for improving the quality of the data and enriches it with additional information. Along with these benefits the development of an automatic or semi-automatic household linking procedure will significantly relieve social scientists from the tedious task of manually linking individuals, families, and households and will therefore improve their productivity. This will allow them to concentrate their time and efforts on the actual analytic research and writing-up of results.

Household linking is different from record linking in several aspects. Traditional record linking compares record pairs of individuals where the similarities of key characteristics remain reasonably stable over time. Household linkage on the other hand seeks to compare pairs of households in which some or even several of the characteristics may change from one census to the next. This suggests that household linkage needs to use richer information than record linking. The emphasis on similarities between record pairs in traditional record linking arises from the fact that a high similarity suggests a good chance of matching two records. Historical census data, however, do not fit this paradigm particularly well. The data they contain are notoriously faulty and, because people's characteristics change across time, i.e., they move house, leave home, marry (and perhaps change name) and change occupations, families and households can change considerably from one census to the next. Adding to these problems is the frequency of common given names and common surnames. Moreover, because record linking is normally used as an interim step towards household linkage, the compu-

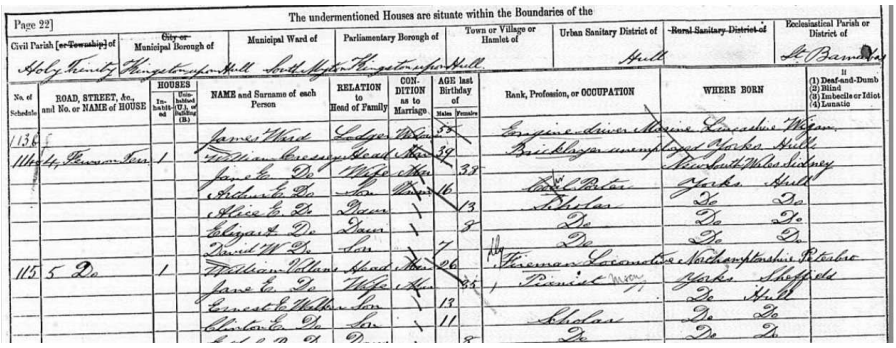

Figure 1: A sample of an original census form.

tation complexity of household linkage is higher than for individual record linkage. Together, these problems not only make it hard to find good matching record pairs, when links are made, many can have the same similarity scores, so that one record in one dataset may be linked to multiple records in another dataset.

Up to now, most research in historical census record linkage has been done by social scientists (Bloothooft 1995, 1998, Fure 2000, Quass & Starkey 2003, Ruggles 2006, Reid et al. 2006, Glasson et al. 2008). Only limited work has used the latest development of record linkage techniques to solve this problem. Vick & Huynh (2011) used the Febrl record linkage system (Churches et al. 2002, Christen & Belacic 2005) to standardise name strings in a population study of census data from the United States and Norway[1]. The authors used name dictionary and statistics of name frequencies to select the names to be cleaned and standardised. Then the Jaro-Winkler approximate string comparison algorithm (Winkler 2006) was used to match candidate names to their standard form. The effectiveness of the standardisation was validated by the fact that it can greatly reduce the number of false links. Goeken et al. (2011) have developed methods to modify the initial record linking results by consideration of the inaccuracy of historical census data collected in the late 19th century. After the initial linkage results were generated by classification of name and age similarity scores using a Support Vector Machine (SVM), name commonness and birthplace density measures were used to generate a set of new linkage results. Weights for each attribute were then generated based on a race, nativity and birthplace analysis on the two sets of linkage results, which lead to the final linked datasets. Larsen & Rubin (2001) looked at the record linking problem from a probabilistic point of view. A mixture model was first selected to divide record pairs into possible matches and non-matches using a maximum likelihood estimation. Then a manual check was performed on the data to update the estimation model. This process was iterated until few additional matches were found. It should be noted that all these work have focused on record linking, but not on household linkage.

In this paper, we introduce a method to link historical census households across time. The major contribution of our approach is to combine supervised learning and group linking methods for household linking. The proposed method first cleans and standardises the census data. Then attribute similarities between pairs of records are calculated. These similarity scores are used as inputs to an SVM classifier, which classifies record pairs into matches and non-matches. Finally, a group linking method is used to match households from different census datasets based on the outcome of the record linking step.

The rest of this paper is organised as follows. Section 2 introduces related work in the areas of data cleaning and record linkage. Section 3 describes the historical census datasets used in this study. A detailed description of the proposed method is given in Section 4, followed by experiments in Section 5. Finally, we draw our conclusions and point out future research directions in Section 6.

2 Related work

In recent years, computer science researchers, mainly in the fields of machine learning, data mining and database systems, have developed new record linkage techniques that can be used to meet the challenges posed by linking historical census data (Kalashnikov & Mehrotra 2006, Bhattacharya & Getoor 2007, On et al. 2007, Herschel & Naumann 2008, Christen 2008b). One recent set of developments are the so called "collective entity resolution" (or collective linkage) techniques (Bhattacharya & Getoor 2007). These techniques use information that explicitly connects records to collectively compute all links between records from two datasets in an overall optimal fashion. The techniques are based on unsupervised machine learning, or use graph-based approaches (Kalashnikov & Mehrotra 2006, Herschel & Naumann 2008). Experimental studies (mostly on bibliographic data) have shown that these techniques can improve linkage quality significantly compared to traditional approaches that consider only pairwise similarities between individual records.

Supervised learning has been investigated for record linking for many years. It uses a training set (labelled examples) to learn a classification model, and then applies the model to testing sets (unlabelled examples) in order to predict the classes of unlabelled examples. Among the supervised learning methods, decision trees and SVMs have been used in record linking (Elmagarmid et al. 2007). The SVM classification technique was developed by Vapnik (1995). It aims at computing a hyper-plane to classify data mapped into a high dimensional space via a kernel function. A key point here is to construct the kernel matrix for which an SVM can be used to perform the training and classification. Bilenko & Mooney (2003) proposed such a solution to compute the similarity of strings and used them as kernel matrix directly. Alternatively, Christen (2008a) constructed inputs to the SVM using a pre-selection step. In this work, a threshold method or nearest-based method was used

[1] Minnesota Population Center: http://www.ipums.org/

	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	6,379	6,848

Table 1: Number of records and households in the UK historical census datasets.

to select record pairs with high confidence of being a match or a non-match. Then these pairs become the positive and negative training samples for the SVM classifier. This method can be considered as a combination of supervised and un-supervised methods.

3 Application Background

The targets of this research are six census datasets collected in ten-year intervals between 1851 to 1901 for the district of Rawtenstall, a small cotton textile manufacturing town in North-East Lancashire. The data were collected on hand-filled census forms, which contains twelve attributes, such as the address of the household, full names, exact ages, sexes, their relationship to the household, occupations and places of birth of each individual residing in his or her accommodation[2]. The hand-filled census forms were transcribed manually onto enumerator's returns sheets. These sheets were subsequently scanned into digital form and, since the late 1990s, various organisations began transcribing the data from these images into tabular form and stored them in spreadsheets where they could be examined by members of the public. A sample of a scanned image is shown in Figure 1. In Table 1, we show the number of records and households in each dataset used on our experiments.

Errors are very common in the transcribed spreadsheets. This is because the original census forms were hand-filled. The English handwriting in the 19th century is quite different from nowadays. The education level of people was low, so even when instructions on how to fill-in the census had been given, many people made mistakes. Enumerators introduced errors when they transferred the data into their enumerator's returns. The quality of the digitisation varies a lot, which was highly related to the personality of the operators and even their gender.

Besides data quality problems, limited and non-standard information in historical census data is another obstacle. The UK 1851-1901 census data contain only twelve attributes (fields) for each record. Many of these attributes change significantly in a ten years interval, such as occupation and geographic mobility. Some attributes do not have values or lack standard values, for example, different names were used for the same occupation. Many nicknames had been used, for example, 'James' is the same as 'Jim', 'Charles' is the same as 'Chas'.

Because of the above problems, reconstruction of family and household data across time is difficult. Social scientists have attempted to clean and link the records manually, but the process is very expensive in terms of time and human resources required. The high cost of cleaning the data and of linking records from one census to another continues to be the principal restriction on their use for academic research.

4 Proposed Method

In this section we provide a detailed description of our proposed approach to household linking, with a focus on the linkage steps of the approach.

[2] www.uk1851census.com

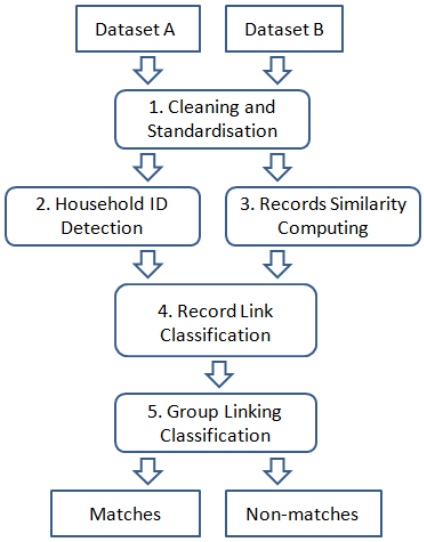

Figure 2: A flowchart of the proposed method.

4.1 Method Overview

The proposed linking method comprises five steps, as is illustrated in Figure 2. The inputs to the system are two datasets to be linked, and the output are record and household pairs that have been classified as matches or non-matches.

The first step in the approach is data cleaning and standardisation. Here, we follow the method proposed by Christen (2008b). The cleaning step aims at eliminating the errors and missing values in the data. It uses look-up tables to remove records without meaningful values, and to replace erroneous attribute values with correct values. An example is the cleaning of gender values, for example, value 'ff' is replaced with 'f'. The standardisation step formats the data into a unified form. It includes several operations, for example, converting values into lowercase letters, splitting first and middle name into two attributes, and unifying the age format into digits-only.

The second step is household detection. The purpose of this step is to assign a unique Household ID (HID) to each household. In the census datasets, we assume that the value for the 'relationship to head' attribute for each household begins with the head of the household. Therefore, we have developed a linear searching algorithm to scan through a census data file, seeking for values for the head of the household, which are 'head', 'head of family', 'widow', 'widower', 'husband', and 'married'. Each time a record has a head of household role, the HID number is incremented by one, and this HID number is assigned to all following records until another record with a head of household role is found (Fu et al. 2011).

The third step is to compute a similarity score for each pair of records under comparison. This step uses several measures to compute the similarities between individual attributes. The attribute similarities are

Attribute	Methods
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 2: Comparison methods used for the six attributes under consideration (Christen 2008b)

concatenated into a vector which is then used in the following classification step. In the last two steps, a record link classification is performed using an SVM, and a group linking classification is used to further improve the linking results.

In the following sections, we will focus on the last three steps of the proposed method. We will address the problem of lacking a ground truth for supervised learning, and how this is solved. We will also show that due to the characteristics of historical census data, domain knowledge can be used to improve both the efficiency and the accuracy of the linking performance.

4.2 Calculating Similarities between Records

We have calculated the similarity for six selected attributes using Febrl (Christen 2008b). Appropriate similarity measures have been chosen for each attribute. A summary of the attributes compared and the corresponding similarity measures is given in Table 2. Details of these measures and their implementation have been described by Cohen et al. (2003) and Christen (2008b).

Here, we give a formal definition of the notion used for our method. Let H_1^i be the i^{th} household in the first census dataset $\mathbf{C}_1$, and $r_i^a \in H_1^i$ be the a^{th} record in this household, with $m_{1,i} = |H_1^i|$ the number of records in household H_1^i, and $1 \leq a \leq m_{1,i}$. Similarly, let H_2^j be the j^{th} household in the second census dataset $\mathbf{C}_2$, and $r_j^b \in H_2^j$ be the b^{th} record in this household, with $m_{2,j} = |H_2^j|$ the number of records in household H_2^j, and $1 \leq b \leq m_{2,j}$.

By concatenating the similarity score calculated for the six attributes shown in Table 2, we get a vector $\mathbf{x}_{r_i^a, r_j^b}$ for record r_a^i from one census dataset and r_j^b from another dataset. For convenience, we denote the similarity vector as $\mathbf{x}_{a,b}$. By summing over the similarity scores, we get a total similarity score $s_{a,b}$. In Table 3, we show the distribution of $s_{a,b}$ on all six historical census datasets under study.

Generally, $s_{a,b}$ reflects the similarity between two records. The larger the similarity value, the more similar two records are. Therefore, a simple way of finding matched pairs of records is to compare the similarity $s_{a,b}$ against a predefined threshold ρ, which is also adopted by the group linkage method by On et al. (2007). If $s_{a,b} > \rho$, the record pair is considered to be a match, otherwise it is considered a non-match. However, there are two problems with this simple method which prohibit effective record linking. Firstly, a number of factors may reduce the total similarity score between two records that belong to the same person. Such factors include, but are not limited to, errors in the data, changes of addresses or surnames, and so on. Therefore, it is difficult to find an optimal ρ for this binary classification scenario. Secondly, the summed similarity score $s_{a,b}$ does not explicitly characterise the contribution of each attribute. In order to take the advantage of separability of all attributes, we should use the full similarity vector, $\mathbf{x}_{a,b}$.

4.3 Classifying Linked Record Pair

To solve the problems with the above simple thresholding method for record linking, we investigated a supervised learning approach. More specifically, we used an SVM to classify the vectors $\mathbf{x}_{a,b}$ obtained from the record pair comparison step.

Training an SVM classifier requires training samples. Because the datasets we obtained do not contain the ground truth in the form of labels of which record pairs are matches or non-matches, we have manually identified 408 true matching record pairs by randomly sampling record links from the 1871 and 1881 datasets. We chose these two datasets because they are the middle ones among the six datasets in our collection. Thus, we assume the sampled pairs have a similar distribution as record pairs sampled from the other pairs of datasets. The labelling process was done as follows. Once a record pair is sampled, we manually decided whether or not the two records are matched. This approach of only labelling record pairs that are clearly matches or non-matches results in training data of high quality which will provide us with an accurate and robust SVM classifier.

Domain knowledge tells us that one record in a dataset must not match with more than one record in another dataset. Therefore, once a record pair is labelled as matched, all other links to the first record in the pair become non-matched. Such a sampling method has generated a large number of non-matched training samples because in the record pair comparison step an exhaustive number of record pairs has been acquired. As a consequence, we have generated a very imbalanced training set, with $314,437$ negative training samples, but only 408 positive samples.

Given the labelled binary dataset $(X, Y) = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, N, y_i \in \{-1, 1\}\}$ (with class 1 being matches and class -1 being non-matches), where $\mathbf{x}_i$ are the indexed similarity vectors $\mathbf{x}_{a,b}$ and y_i are their labels, an SVM classifier recovers an optimal separating hyper-plane $\mathbf{w}^T\mathbf{x} + b = 0$ which maximises the margin of the classifier. This can be formulated as the following constrained optimisation problem (Vapnik 1995):

$$\min_{\mathbf{w},b,\xi} \quad \frac{||\mathbf{w}||^2}{2} + C\sum_i \xi_i \qquad (1)$$

$$\text{s.t.} \qquad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \leq 1 - \xi_i \text{ and } \xi_i \leq 0$$

Here, a function ϕ is used to map the training vectors $\mathbf{x}_i$ into a higher dimensional space. $C > 0$ is the penalty parameter of the error term, and ξ is the margin slack variable. To handle the situation of imbalanced training data, we can assign a large penalty parameter for the positive class and a much smaller one for the negative class. In this study, we have set C^+ to $1000 \times C^-$.

After training, the SVM classifier is used to classify all record pairs generated by pair-wise linking all six datasets. In Table 4, we show the results of the total number of record pairs that are classified as matches, and the statistics of the number of records with single and multiple matches. As mentioned before, a record in a dataset should only be matched to at most one record in another dataset. Therefore, we have to remove those multiple matches.

	1851-1861	1861-1871	1871-1881	1881-1891	1891-1901
$s_{a,b} \in [0,1)$	431,610	705,570	891,011	981,225	1,048,323
$s_{a,b} \in [1,2)$	2,101,264	2,760,774	3,277,425	3,332,895	3,211,875
$s_{a,b} \in [2,3)$	1,926,086	2,437,898	2,860,517	2,865,787	2,665,369
$s_{a,b} \in [3,4)$	591,115	724,945	824,939	857,084	831,405
$s_{a,b} \in [4,5)$	55,053	64,462	65,908	62,317	60,316
$s_{a,b} \in [5,6)$	2,721	3,826	4,160	3,865	4,837
$s_{a,b} = 6$	187	278	239	76	296

Table 3: Distribution of Similarity scores $s_{a,b}$ on six historical census datasets.

	1851-1861	1861-1871	1871-1881	1881-1891	1891-1901
Total matched record pairs	56,301	71,752	80,802	80,504	79,442
Records involved in a single match	3,782	5,059	6,818	7,748	7,946
Records involved in multiple matches	8,784	10,910	11,965	10,406	13,034

Table 4: Record linking results on six historical census datasets based on SVM classification.

4.4 Group Linking

The group linking step aims at linking households based on the classified record links. Because the number of matched pairs generated by the SVM are larger than the number of records in both datasets, there are many multiple links. In the group linking step, if the households of the matched records in the multiple links will be compared against H_1^i one by one, then unnecessary household linking will be performed, which makes the step not efficient.

To solve this problem, three strategies can be adopted. Firstly, we can remove multiple record links by simply choosing the matched pairs with the highest $s_{i,j}$ values for each r_i^a. This will generate either a unique record link, or multiple but less record links when several links have the same $s_{i,j}$ score for each r_i^a. However, as we mentioned previously, due to erroneous data or changes in the data, exact matches are difficult to find, and $s_{i,j}$ may be low. Therefore, a true record match may not be at the top when ranked using $s_{i,j}$ only, and such a strategy will remove many true matches. The second method is to set a threshold ρ to help the decision. Record links with $s_{i,j} < \rho$ can be removed from consideration. Even if such a threshold is set, one record in a dataset still can be linked to several records in another dataset, because the corresponding similarity scores are too close or identical. Alternatively, as a third method, we can keep all record links in the group matching step. Because several linked records may belong to the same household, we calculate the best unique pairs of households that match across two census datasets.

Several group linking techniques have been proposed for bibliographic record linkage (Bhattacharya & Getoor 2007, Herschel & Naumann 2008, Kalashnikov & Mehrotra 2006). In this research, we modify the method by On et al. (2007) to link two households. For each pair of linked households, the household similarity score $\mathbb{S}_{i,j}$ between two households, H_1^i and H_2^j, can be calculated using the normalised weight of the matched individual record pairs in the two households:

$$\mathbb{S}_{i,j} = \frac{\sum_{(r_i^a, r_j^b) \in M} sim(r_i^a, r_j^b)}{m_{1,i} + m_{2,j} - |M|}. \quad (2)$$

where M is the set of record pairs matched between H_1^i and H_2^j. Here the similarity function $sim(r_i^a, r_j^b)$ can take two forms:

$$sim(r_i^a, r_j^b) = 1, \quad (3)$$

for taking the labels of matched record pairs predicted by the SVM, or

$$sim(r_i^a, r_j^b) = s_{i,j}, \quad (4)$$

for taking the sum of the attribute-wise similarity. In the former case, the group linking reduces to computing the Jaccard index (Tan et al. 2005). The second form corresponds to solving a weighted bipartite matching problem (Chartrand 1985).

Matched households can be classified by selecting the links with the highest $\mathbb{S}_{i,j}$ value. Here we assume that a household in one dataset can be matched to at most one household in another dataset. It should be mentioned here that this assumption does not always hold. The children in a household may get married or move out during the interval between two censuses. Therefore, a household can split into multiple households. However, as we mentioned at the beginning of the paper, the purpose of household linkage is to find the households which have a majority of their members matched. Thus, our purpose is to link the most 'stable' part of households.

We summarise our group linking approach in Algorithm 1. The input to the algorithm are all the matched record pairs Π between the two datasets $\mathbf{C}_1$ and $\mathbf{C}_2$, and a household $H_1^i \in \mathbf{C}_1$. The output is the household $H_2^{j*} \in \mathbf{C}_2$ which has the highest similarity to H_1^i. From Π, we can find all records in $\mathbf{C}_2$ that match to records in household H_1^i. Each of these matched records belongs to a household in $\mathbf{C}_2$, and some of them might belong to the same household. To improve the efficiency of household matching, we then merge duplicate households, so that only unique households will be used to calculate the similarities to H_1^i using Equation 2. Finally, the household(s) with the highest similarity $\tilde{\mathbb{S}}_{i,j}$ will be selected as the output H_2^{j*}.

Step 4 in Algorithm 1 is important because it improves the efficiency of the proposed method. This is because several records in a household may be matched to other records that belong to the same household. Therefore, finding unique households will reduce the number of household similarity calculations. An example of this situation is shown in Figure 4. The four records in household A are matched to five records in households B and C. Instead of calculating household similarities five times, by finding the unique matched households, we only need to conduct two similarity calculations. In this case, the number of household pairs to be linked is two.

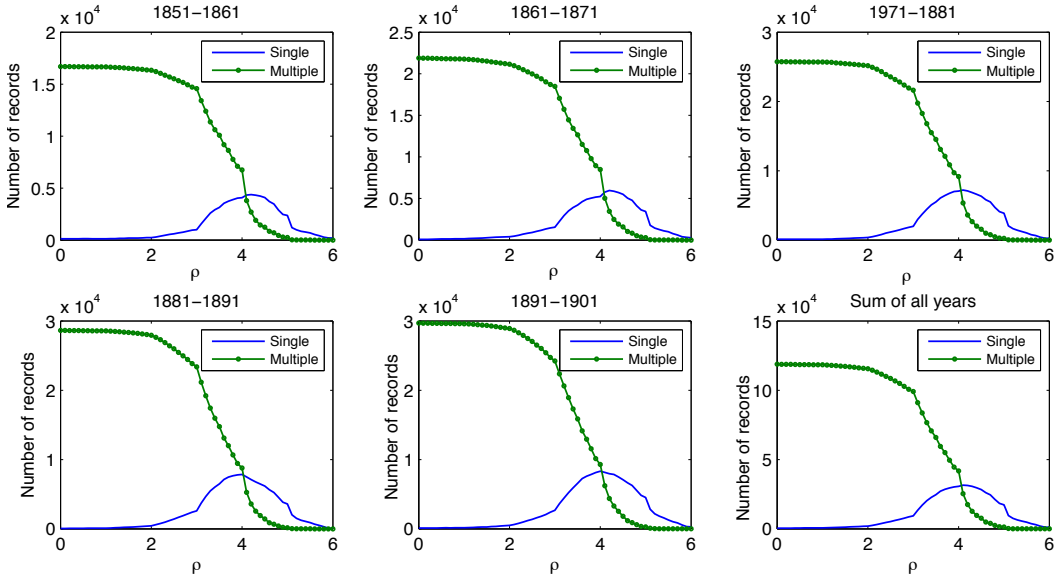

Figure 3: Record linking results using the thresholding method.

Algorithm 1: Group Linking

Input:
- Matched record pairs:Π
- All households in the second dataset: $\mathbf{C}_2$
- A household in the first dataset: H_1^i

Output:
- Best matching household: H_2^{j*}

1: **for** $r_i^a \in H_1^i$ **do**
2: Find all matched records $\{r_j^b\} \subset \mathbf{C}_2$ inΠ
3: Find households $\{H_2^j\} \subset \mathbf{C}_2$ for all r_j^b
4: Find unique households $\{\tilde{H}_2^j\} \subseteq \{H_2^j\}$
5: Calculate household similarities $\{\tilde{\mathbb{S}}_{i,j}\}$ for H_1^i and $\{\tilde{H}_2^j\}$ using Equation 2
6: Find H_2^{j*} with maximum $\tilde{\mathbb{S}}_{i,j}$

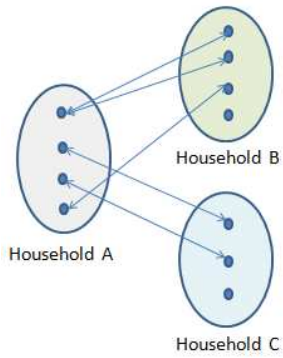

Figure 4: Example of the household (group) linking approach.

5 Experiments and Evaluation

We have conducted experiments on all six historical census datasets following the steps introduced in the previous sections. We used LIBSVM (Chang & Lin 2011) with an RBF kernel for training and testing of the record pair similarity vectors. To cope with the extremely unbalanced data in the training set, we have set the penalty parameter for the positive class to be $C^+ = 1000$ and for the negative class to be $C^- = 1$.

5.1 Experiments on Record Linking

First, we compare the performance of the SVM classifier against the thresholding method for the record linking step. Let's first consider the thresholding method. The similarity score $s_{a,b}$ for each pair of records r_i^a and r_j^b can be calculated by adding all attribute scores together. Appropriate setting of the thresholding parameter ρ is important when separating record pairs r_i^a and r_j^b into the matched and non-matched classes. We solve this problem by analysing the linking results with respect to the value of ρ. Figure 3 shows the number of records in one dataset with exactly one matched record and the number with multiple matched records in the other dataset, when different values for ρ have been set. The distribution of single matched records and multiple matched records are different for different ρ. Increasing ρ can reduce the number of records with multiple matches.

From Figure 3, two further observations can be obtained. Firstly, the curves in each plot follow a similar trend. This is consistent with the distribution of similarity scores shown in Table 3. This observation is important, because it suggests that a model trained on record similarities from any pair of datasets, or tuned on these datasets, can be applied directly to classify record pairs in other pairs of datasets as well. Secondly, the curves for only one match and for multiple matches intercept at certain points. We claim that these points can be set as the default ρ value for the group linking step. Therefore, we set $\rho = 4$ for the linking of all pairs of census datasets.

Using an SVM to perform classification of record pairs is more straightforward. As mentioned in the previous sections, we manually labelled some matched pairs in the 1871 and 1881 datasets, in total 314,437 training samples. We trained an SVM using this training set. After that, we used the trained model to classify record pairs into matched and non-matched classes, which generated the results in Table 4.

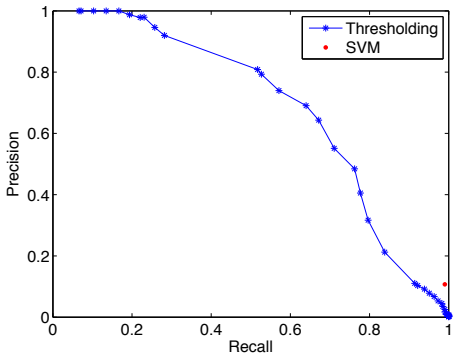

Figure 5: Training set precision and recall for SVM and thresholding for different ρ values.

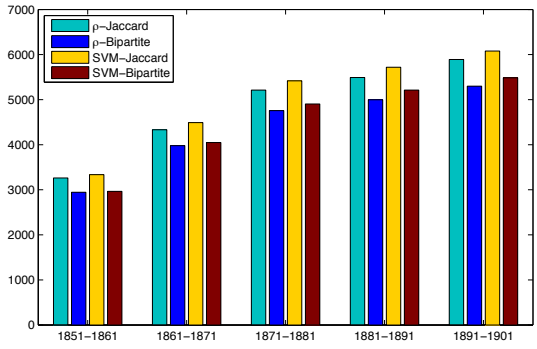

Figure 6: The number of households matched with different methods for the group linking step.

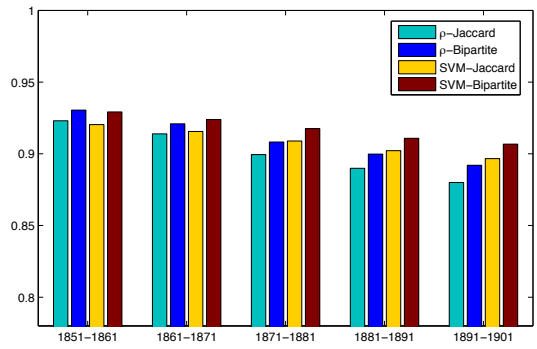

Figure 7: Group linking results shown as the percentage of reduction in the number of matched households.

We used the training set to compare the performance of the thresholding method and the SVM classification. We found that many true links had been missed when ρ was set too high in the thresholding method. For example, when ρ was set to 5.5, only 80 out of 408 pairs of matched records were found and there were no multiple matches. On the other hand, when ρ was too low, many multiple matches were generated. When ρ is set to 4, as suggested previously, 3,384 pairs of matched records were found, including 373 true matches. The SVM has generated 3,371 pairs of matched record with 404 true matches.

For further comparison, in Figure 5, we show the precision-recall curve when ρ changes. The precision and recall of the SVM classification is plotted using a red dot on the lower-right side of the graph. This plot suggests that at the same recall level, the SVM classification generates better precision than the thresholding method. The high recall score of the SVM guarantees that most true matches are retrieved. Though a high number of false matching record will be generated, this number can be greatly reduced in the following group linking step.

5.2 Group Linking

With the record linking results ready, we can perform the group linking step. Here we would like to compare four combinations of record linking and group linking methods. The methods for record linking include thresholding with ρ, and SVM classification. The methods for group linking include using either Jaccard or Bipartite metrics for the group similarity calculation. We label these four methods as ρ-Jaccard, ρ-Bipartite, SVM-Jaccard, and SVM-Bipartite. Here, we have set $\rho = 4$ for all experiments.

We start by showing in Table 5 the number of matched record pairs after the thresholding and SVM classification steps. For each of these pairs, the households that contain the record pair should also be compared. As described in Algorithm 1 and Figure 4, the number of household links can be reduced by finding the best unique household to be linked. In Table 5, we also show the number of households to be linked after such optimisation. It can be seen that the number of households generated by the SVM classification is higher than those generated by the thresholding method. This is because the number of matched record pairs for the former is higher than those from the latter. As mentioned earlier, the SVM classification generates a high number of matching records. This guarantees that less households are missed in the matching process. As a consequence, among the households detected in this step, there are many multiple links that have the same group similarity score, so that a household in the first dataset may be matched to multiple households in the second dataset.

After the group linking step, the number of matched households is greatly reduced. The total number of matched households for each matching period is shown in Figure 6, while the percentage of reduction is shown in Figure 7. From Figure 6, we can observe that when using bipartite matching in the group linking, the number of matched households is lower than the Jaccard index counterpart. This suggests that the bipartite matching is more powerful in removing multiple matches. We can also observe that the SVM-based methods generate more matched households than the thresholding-based methods, except for the period of 1851-1861. This is due to the fact that the record matching step has generated more matched record pairs when SVM classification is applied than for the thresholding method.

Figure 7 shows that higher reduction rates have been achieved on the SVM-based methods compared to the thresholding methods proposed by On et al. (2007). This is especially the case for the census datasets after 1871. In fact, all four methods under comparison have achieved high reduction rates of multiple links, with more than 87% multiple matched households removed in all the periods.

To further analyse the composition of matched households, in Table 6 we report statistics on households with single and multiple matches for the four methods under comparison. As can be seen, the num-

	1851-1861	1861-1871	1871-1881	1881-1891	1891-1901
Matched record pairs by thresholding	57,961	68,566	70,307	66,312	65,449
Household pairs to be linked after thresholding	42,360	50,312	51,815	49,868	49,070
Matched record pairs by SVM	56,301	71,752	80,802	80,504	79,442
Household pairs to be linked after SVM	41,900	53,214	59,473	58,435	58,816

Table 5: Number of matched records and household comparisons to be performed after the record linking step.

ber of households with a single match is much larger than the number with multiple matches. This suggests that our group linking method is very effective in removing the multiple matches generated in the record matching step. Among all four methods, the SVM-Bipartite method has achieved the highest number of single matches, as well as the lowest number of multiple matches. This has made this method suitable for application to historical census household linkage.

Finally, we show in Table 7 the number of households in the 1851 dataset that have been linked in periods of different lengths. The linking used the group linking results for each 10 year period reported above. For a household in the 1851 dataset, we first identified its match(es) in the 1861 dataset, then the match(es) in the 1871 dataset for each matched household in the 1861 dataset. The process continues iteratively until no match(es) can be found or until we have gone through all the datasets. All four methods have detected more than 2,200 households that have been linked over a period for 50 years. Only less than 200 households have disappeared every 10 years. Such results may occur for two reasons.

Firstly, the group linking is based on the record linking step. As long as record matches can be found for a member in a household for a 10 year period, the household linking continues for the next 10 year period. This means even if members in a household have perished or moved away, the linking process can be continued if at least one household member can be found in the following census datasets. The fact that a large number of household links has been found for the whole 50 year period tells that some children in a household tended to stay in the same area as their parents even when they've grown up and formed a new family. Therefore, such a process has generated the possibility of tracing family trees. We will manually evaluate these results with domain experts.

Secondly, such results may also be due to false matches in the record linking step. Although it is hard to judge the correctness of such matches due to lack of ground truth information, this study provides social scientists with a means to trace household changes across time. As far as we know, this is the first work of this kind in the field of historical census record linkage.

6 Conclusion

In this paper, we have introduced a novel approach to historical census household linkage. This approach first computes the similarity between record pairs. Then these similarities are used as input to an SVM classifier, which classifies record pairs into a matched and a non-matched class. The classification outcome forms the input to the household linking step. We have used a group linking technique to generate household linking similarities. The Jaccard and Bipartite measures are used in the group linking models, and their performance is compared. The results show that when combining support vector machine classification for record linking with group linking using bipartite matching, the household linkage generates better results than the alternative methods under comparison. This paper shows that the combination of supervised learning and group linkage methods for historical census household linkage is very effective. It provides social scientist with novel tools to analyse historical census data.

In the future, we will explore interactive and iterative learning methods to improve the supervised learning model. This includes learning from the instances where a household has split into multiple households between two censuses, and exploring other supervised learning approach as solution. We also plan to use a forward and backward linking method to further improve the household linking process over 20 to 50 years periods, and have the results evaluated by domain experts.

References

Anderson, M. (1971), *Family structure in nineteenth century Lancashire*, Cambridge: Cambridge University Press.

Bhattacharya, I. & Getoor, L. (2007), 'Collective entity resolution in relational data', *ACM Transactions on Knowledge Discovery from Data*, **1**(1).

Bilenko, M. & Mooney, R.J. (2003), Adaptive duplicate detection using learnable string similarity measures, *in* '9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 39–48.

Bloothooft, G. (1995), 'Multi-source family reconstruction', *History and Computing*, **7**, 90–103.

Bloothooft, G. (1998), 'Assessment of systems for nominal retrieval and historical record linkage', *Computers and the Humanities*, **32**(1), pp. 39–56.

Chang, C.-C. & Lin, C.-J., (2011), 'LIBSVM: A library for Support Vector Machines', *ACM Transactions on Intelligent Systems and Technology*, **2**(3), pp. 27.

Chartrand, G., (1995), *Introductory Graph Theory*, New York: Dover.

Christen, P. & Belacic, D. (2005), Automated probabilistic address standardisation and verification, *in* 'Australasian Data Mining Conference', pp. 53–68.

Christen, P. (2008a), Automatic training example selection for scalable unsupervised record linkage, *in* '12th Pacific-Asia Conference on Knowledge Discovery and Data Mining', Osaka, pp. 511–518.

Christen, P. (2008b), Febrl: An open source data cleaning, deduplication and record linkage system with a graphical user interface, *in* '14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 1065–1068.

	1851-1861		1861-1871		1871-1881		1881-1891		1891-1901	
	S	M	S	M	S	M	S	M	S	M
ρ-Jaccard	2,642	272	3,624	309	4,326	384	4,578	395	4,845	442
ρ-Bipartite	2,889	25	3,896	37	4,671	39	4,951	22	5,275	12
SVM-Jaccard	2,668	293	3685	357	4,497	398	4,805	405	5,025	456
SVM-Bipartite	2,956	5	4,035	7	4,886	9	5,208	2	5,478	3

Table 6: Number of households identified with single (S) and multiple (M) matches for all linked datasets.

	10 Years	20 Years	30 Years	40 Years	50 Years
ρ-Jaccard	134	133	136	119	2,392
ρ-Bipartite	140	158	165	156	2,295
SVM-Jaccard	121	132	134	132	2,442
SVM-Bipartite	120	147	157	146	2,391

Table 7: Households linked in time periods with different lengths.

Churches, T., Christen, P., Lim, K. & Zhu, J.X. (2002), Preparation of name and address data for record linkage using hidden Markov models, 'BMC Medical Informatics and Decision Making', Vol. 2, no. 9.

Cohen, W.W., Ravikumar, P. & Fienberg, S.E. (2003), A comparison of string distance metrics for name-matching tasks, *in* 'IJCAI-03 Workshop on Information Integration', pp. 73–78.

Elmagarmid, A.K., Ipeirotis, P.G. & Verykios, V. S. (2007), 'Duplicate Record Detection: A Survey', *IEEE Transactions on Knowledge and Data Engineering*, **19**, 1, 1–16.

Fu, Z., Christen, P. & Boot, M. (2011), Automatic cleaning and linking of historical census data using household information, *in* 'Workshop on Domain Driven Data Mining, held at IEEE ICDM'11', Vancouver.

Fure, E. (2000), 'Interactive record linkage: The cumulative construction of life courses', *Demographic Research*, **3**, 11.

Glasson, E., De Klerk, N., Bass, A., Rosman, D., Palmer, L. & Holman, C. (2008), 'Cohort profile: The western Australian family connections genealogical project', *International Journal of Epidemiology*, **37**, 30–35.

Goeken, R., Huynh, L., Lynch, T.A. & Vick, R. (2011), 'New methods of census record linking', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **44**(1), 7–14.

Herschel, M. & Naumann, F. (2008), Scaling up duplicate detection in graph data, *in* '17th ACM Conference on Information and Knowledge Management', pp. 1325–1326.

Kalashnikov, D.V. & Mehrotra, S. (2006), 'Domain-independent data cleaning via analysis of entity-relationship graph', *ACM Transactions on Database Systems*, **31**(2), 716–767.

Larsen, M.D. & Rubin, D.B. (2001), 'Iterative automated record linkage using mixture models', *American Statistical Association*, **79**, 32–41.

On, B.W., Koudas, N., Lee, D. & Srivastava, D. (2007), Group linkage, *in* 'IEEE 23rd International Conference on Data Engineering', pp. 496–505.

Quass, D. & Starkey, P. (2003), Record linkage for genealogical databases, *in* '2003 ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation', pp. 40–42.

Rabiner, L. R. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE*, **77**(2), 257–286.

Reid, A. , Davies, R. & Garrett, E. (2006), 'Nineteenth century Scottish demography from linked censuses and civil registers: a "sets of related individuals" approach', *History and Computing*, **14**(1+2), 61–86.

Ruggles, S. (2006), 'Linking historical censuses: A new approach', *History and Computing*, **14**(1+2), 213–224.

Tan, P., Steinbach M. & Kumar V. (2005), *Introduction to Data Mining*, Pearson Addison-Wesley.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag.

Vick, R. & Huynh, L. (2011), '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**(1), 15–24.

Winkler, W. E. (2006), Overview of research linkage and current research directions, US Bureau of the Census, Statistical Research Report Series RRS2006/02.