Record Linkage Techniques for Improving Online Genealogical Research using Census Index Records

John S. Lawson
Department of Statistics, Brigham Young University Provo, UT 84602, lawson@byu.edu

Abstract

Since the first edition of Alex Haley's book, *Roots*, in 1976, there has been a fascination with tracing family origins. A Google search for "genealogy" results in more than 26 million websites. Many sites such as Ancestry.com and HeritageQuest.com offer users online searches of census index records, which are electronic databases containing the basic information for each head of household from the census records. However, unless the user enters the same name, spelling, and age that is recorded in the census index records, the online search could be very frustrating. This paper describes the implementation of Probabilistic Record Linkage to improve the search. By sampling the 1910 and 1920 census index records, record linkage weights were developed. Using these weights, the probability of missing the record sought was reduced to an estimated 1.23%, with a reasonable number of false hits.

Keywords: EM Algorithm, Online Search, Family History, Probabilistic Record Linkage, Multinomial Distribution,

1. Introduction

Interest in family history research that was generated by Alex Haley's book, *Roots*, has not waned in the thirty years since he published it. In fact, interest in genealogy has increased. According to a recent national telephone survey (Maritz, 2000), the percentage of the U.S. population that is interested in family history research or genealogy has increased from 45% in 1996 to 60% in 2000, and 35 million people have used the internet for online family history research.

Many online resources exist for those interested in tracing their family origins. Cyndi's List www.cyndislist.com/database.htm includes web sites with databases of information such as vital records, land records, census records and military records. A good place to start looking for an ancestor is in the census records. If you know the state and county where your ancestors lived during one of the decennial census years, and you have the names and approximate ages, you might locate them in the census index records online.

Both Ancestry.com and HeritageQuest.com allow users to search census index records online. Census index records contain a summary of the information on each head of household and all other members of the household with a different last name from the full census records. Census index records are electronic records that can be searched by entering your information into an online search form. Once a record that might represent your ancestor is identified in the census index records, you can click on a link that will display a photocopy of the complete original handwritten census page from which the summary information was extracted. The complete census record gives other useful information like the names and ages of other members of the household, names of parents, place of birth, occupation, naturalization status, and street address, etc.

The census index records contain nine fields: surname, given name, age at the time of the census, gender, race, birthplace, state, county and locale or census district. At least one of the online search programs allows you to enter information on each of these fields into the search form. If you have information for all of these fields, you can enter each of them. If you don't know some of the entries, such as the age or county of residence, you can leave those fields blank and the search program treats them like a wild card.

The problem with the currently available search programs is that name spelling, reported ages, birthplace, and race are not always correct or consistent from census to census, so the spellings and reported ages may be slightly different than the information you have at hand. When you enter your information into the search form, the program tries to find a record in the census index database that matches the fields entered exactly. If there is a difference in one or more of the fields you enter, the program will fail to find the record sought. If you leave too many fields (such as given name and age) blank as wildcards, you could get tens of thousands or even millions of "hits"

(too many to reasonably examine).
If a user locates an ancestor in the 1910 census index records, there is less than a 50% chance that the same person will have identical fields of information recorded in the 1920 census index. For example, I found my grandfather, James Lawson, in the 1910 records, but when I searched for him in the 1920 records with the same input fields, I got the message "no records found" (see Lawson 2006). However, when I manually searched through microfilm records at a Family History Library, I noticed a JS Lawson listed with the same age and birthplace. When looking at the detailed record, I could see this was the same person, except that he had been enumerated by his initials rather than by his first name. A search program that looks for exact matches will not be able to recognize a likely variation that could represent the same individual.

My experience with the computer search of census records is not unlike many novices who go online. When they first hear of the websites where they can find information about an ancestor, they anxiously go to the site, fill out all the boxes in the search form, click the search button and either find nothing or a very long list of information to examine. They don't realize why they were unsuccessful, and they give up in frustration instead of seeking help or reading one of the many books (such as Kemp 2003 or Crume 2004) or articles that describe productive strategies for searching online databases.

A computer search program that looks for exact matches will not be able to decide intelligently when a variation in two fields may represent the same person. Fortunately, there is a solution to this problem. Computer search programs can be programmed to use human-like intelligence when comparing records in a file to the information in the search form using a technique called Probabilistic Record Linkage.

A search program that uses Probabilistic Record Linkage works by comparing each field in a record to the corresponding field in the input form, but rather than rejecting every record that doesn't match exactly on one or more of the fields, the program assigns a weight to each field based on the comparison. If a field on the record, such as given name, matches exactly the same field on the input form, then a positive weight is assigned. If a field on the record is different from the same field on the input form, a negative weight is assigned. If the fields do not match exactly, but are close (for example, if one is an abbreviation or initial of the given name in the comparison field) a weight somewhere between the positive weight given for matching fields and the negative weight given for differing fields is assigned. Once a weight is assigned to each field of a record, a score is calculated by adding the weights across all the fields in a record. Records that have a high proportion of fields matching the input screen will have large positive scores, and records with a low proportion of matching fields will have lower or negative scores. Each record whose score exceeds a threshold value will be displayed in the list of possible matches in the search results.

Three of my students and I have attempted to create a Probabilistic Record Linkage Program for searching census index records. Francis (2004) calculated linkage weights using data from the 1910 and 1920 census index records for Ascension County in Louisiana. This was followed by Jensen (2004) who computed the weights for five additional data sets, described in Section 2, and determined whether an average set of weights could be used on all 1910 and 1920 census index records. Finally, Bauman (2006) investigated the use of the EM Algorithm to calculate weights in order to avoid the time-consuming clerical matching done by Francis (2004) and Jenson (2004) to obtain a training set. A prototype search program has been written and is now in the process of being implemented at HeritageQuest Online.

2. Calculating PRL Weights for a Search of Census Index Records

The weights and threshold values used in a Probabilistic Record Linkage search program should not be determined arbitrarily. Otherwise, the program may miss the record sought or display too many possible "hits" in the search results. The goal is to choose weights that will maximize the chance of finding the record sought and minimize the number of incorrect records displayed in the search results. We calculated the optimal weights for $\nu = 1, \cdots, V \leq 9$ fields in the 1910 and 1920 census index records, where 9 is the total number of fields in the records. I will describe how we did this using the notation of Winglee, Valliant and Scheuren (2005).

If we consider possible pairs of records from the census index records where one record is from the 1910 census index and the other is from the 1920 census index, then each pair of records, *r*, belongs to one of two classes. Either both records in the pair represent different people $(r \in U)$ or both records represent the same person in different census years $(r \in M)$.

When comparing the ν^{th} field between two records in

a pair, there are $i = 1, \cdots, c_v$ mutually exclusive possible outcome events. For example i = 1 could indicate the event that the fields are the identical for the two records; i = 2 might indicate the event that the fields are not the same but similar; and i = 3 might indicate the event that they are totally different. We denote the event indicator vector of the comparison for the v^{th} field on the rth pair of records as $\mathbf{y}_{rv} = (y_{rv1}, \cdots, y_{rvc_v})$, where if $y_{rvi} = 1 \Rightarrow y_{rvj} = 0$, for $i \neq j$ since the outcome events are mutually exclusive.

For a large data file, define m_{vi} to be the conditional probability of obtaining outcome category i when comparing field v between a randomly selected pair of records that actually represent the same individual in two different census years (i.e. $r \in M$). Then the conditional distribution of $\mathbf{y}_{rv}$, given both records in the pair represent the same person, is approximately a unit multinomial with parameters $m_{v1}, \cdots, m_{vc_v}$, i.e.

$$P\{\mathbf{y}_{rv} \mid r \in M\} = \prod_{i=1}^{c_v} m_{vi}^{y_{rvi}} .$$

Similarly, we define u_{vi} to be the conditional probability of obtaining outcome category i when comparing field v between a pair of records randomly selected from those that do not represent the same individual (i.e. $r \in U$). The conditional distribution of $\mathbf{y}_{rv}$, given the two records in the pair represent different people, is a unit multinomial with parameters $u_{v1}, \cdots, u_{vc_v}$.

When a Probabilistic Record Linkage program compares a pair, r, of records, it is essentially testing the hypothesis that the pair of records in question represents two different people, versus the alternative that it represents the same person (i.e., $H_0 : r \in U$ vs $H_a : r \in M$). To do this, Fellegi and Sunter(1969) showed that the optimal weight for the agreement status (i.e., identical, similar or different) of each field was the log odds of obtaining that status with $(r \in M)$ as opposed to $(r \in U)$. This was based on the statistical theory of most powerful tests of a simple hypothesis versus a simple alternative.

Assuming the event indicator vectors from the v fields are independent, and following the logic of the most powerful test of a simple hypothesis versus a simple alternative, the test statistic or score for record pair r (or log likelihood ratio) is $S = \sum_{v=1}^{V} w_v$, and

$$w_v = \sum_{i=1}^{c_v} y_{rvi} [\ln m_{vi} - \ln u_{vi}] \qquad (1)$$

is the optimal weight for the outcome event indicator y_{rv} assigned when comparing the v^{th} field. When this score function exceeds a threshold T (e.g., critical value) for any pair of records r, that pair would be considered to represent the same person. In order to calculate these optimal weights and scores exactly for a set of census index records, the conditional probabilities $m_{v1}, \cdots, m_{vc_v}$ and $u_{v1}, \cdots, u_{vc_v}$ must be known or estimated for each field, v.

To do this we started with a sample of census index records for both the 1910 and 1920 census. The sample included records from 12 counties in five states: Yolo, Del Norte, Lassen, Kings and Ventura counties in California; Tolland county in Connecticut; De Kalb and Hamilton counties in Illinois; Huron, Crawford and Oceana counties in Michigan; and Ascension county in Louisiana. These records were grouped into six data files shown in the first column of Table 1. The second column of Table 1 shows the total number of records from each file (1910 and 1920 census records combined). These records were sorted in several ways and pairs of records that appeared to represent the same person (or a "match" i.e., $(r \in M)$) were determined clerically. This was done separately with the records from each file, and the numbers of "matches" found are shown in the third column of Table 1. If there was any doubt whether two records represented the same person, images of the original census records could be checked to verify the paring.

Table 1 Description of Sample Census Index Records

Data File	Total Number of records from 1910 & 1920	Number of clerical Matches found in both census years	Matches not within the same block
Connecticut	18,799	2,574	169
Illinois	32,211	5,206	222
Michigan	34,497	4,879	340
So. Calif.	32,684	2,914	135
No. Calif.	21,436	2,049	106
Louisiana	14,218	641	45
Totals	153,845	18,263	1,017

Next, using SAS proc sql we created a file, *F*, of pairs of records for each file. We paired records so that records from the 1910 census index were paired with records from the 1920 census index. To reduce the total number of pairs to a workable amount, we first

blocked the records by surname and gender. We only paired 1910 census index records with 1920 census index records within the same surname by gender block, dramatically reducing the number of pairs. For example, in the Louisiana records there were 7537 records from 1910 and 6681 records from 1920, and the total number of possible pairs was $7537 \times 6681 = 50{,}384{,}697$. By blocking, the number of pairs was reduced to 88,471. Using this paring procedure, a number of the clerical "matches" found earlier would not be paired together because there were differences in the surname spelling and/or gender between the 1910 and 1920 census index records. The numbers of "matches" that were not paired together are shown in the fourth column of Table 1.

Let $r \in F$ represent the record pair index for a pair of records in the file *F*, where *r* = 1,…,*R* and *R* represents the total number of record pairs we created. Each pair of records in the file falls into one of two subsets. A pair represents the same person in two different census years, or it represents two different people. Thus, $F = M \cup U$. Each pair of records could be classified into one of these two subsets since the earlier clerical matching had identified all pairs that represented the same person. The proportion of the records in *F* that belonged to the subset *M* was small. For example, in Louisiana the proportion of records in *M* was (641-45)/88,471=0.0067.

Once the pairs of records that represented the same person were determined, the conditional probabilities, $m_{v1}, \cdots, m_{vc_v}$ had to be determined. To do that we first had to define the comparison events for each field. This is a subjective decision and will be different depending on the records being matched and the insight of the person defining the event categories. For our first attempt, we defined four possible outcome categories for comparing given names, three outcome categories for comparing ages, three categories for comparing race, and two outcome categories for comparing birthplace and county. We decided that if the given names matched exactly or if one was a recognized nickname of the other, we would call them the same (i.e., $y_{r1_1} = 1$). If the given names were not the same nor was one a nickname of the other but they matched on the first three letters, we classified the comparison as close (i.e., $y_{r1_2} = 1$). If the names were not the same and didn't match on the first three letters but did match on the first letter, we classified the comparison as a second degree close (i.e., $y_{r1_3} = 1$). Finally, if the comparison of given names resulted in anything other than what we had already defined, we classified the comparison as different (i.e., $y_{r1_4} = 1$). The categories for given name and the other fields we defined are shown as the column headings in Table 2.

In the same SAS proc sql program we used to create the file of pairs, F, we compared the fields between each pair of records and appended the indicator variables $\mathbf{y}_{rv} = (y_{rv1}, \cdots, y_{rvc_v})$ that resulted from the comparisons. Next, we estimated the conditional probabilities $m_{v1}, \cdots, m_{vc_v}$ by counting the relative frequencies of the various outcome events, i.e., $\hat{m}_{vi} = \bar{y}_{\cdot vi}$ for $r \in M$. The frequency counts were made using SAS proc freq. Since most of the pairs of records, *r*, in the file *F* do not represent the same person, the relative frequency of various outcome events obtained from a random sample of pairs of records in $r \in F$ was a reasonable estimate of $u_{v1}, \cdots, u_{vc_v}$, and we used $\hat{u}_{vi} = \bar{y}_{\cdot vi}$ for a random sample as estimates of these conditional probabilities. We selected a random sample of records for the 1910 records and the 1920 records separately using SAS proc surveyselect. We next sorted the samples in a random order and merged them together to form a file of random pairs. We again used SAS proc freq to estimate the conditional probabilities from this file.

From the estimated conditional probabilities the weights in equation (1) were calculated separately for each data file shown in Table 1. In general, a match on given name resulted in the widest range in weights, or most discrimination power, followed by age, birthplace, and county. The race field had the least discrimination. No weights were calculated for surname and sex, since these two fields were used to block the records prior to pairing. No weight was calculated for census local, since classifications were often subjective and many definitions changed from 1910 to 1920.

Between different data files the field weights for the event status "same" were not equal but were quite consistent. The weights for events "close" and "different" were not as consistent from file to file as the event status "same," but were still fairly consistent.

3. Results

3.1 Weights

Since the weights were somewhat consistent for each data set, overall weights were created by averaging the weights calculated in each data file. Initially these

weights were tested by applying them to pairs of records in the file we created by blocking on surname and age. The averaged weights for each field are shown in Table 2.

Table 2 Averaged Weights

Field	Weight for Same	Weight for Close	Weight for Different
Given name	4.18	Nicknm 4.18 / 1st 3 3.29 / 1st 0.36	-4.76
Age	±3 yrs 2.46	±4 yrs -0.11	-2.63
Race	0.18	One B and other M 0.84	-1.59
Birthpl	1.50	Match Lookup Table 1.50	-2.67
County	0.50		-3.16

The program also used two lookup tables. A nickname table was based on the list of common nicknames from the USGenWeb Project[1999/2000] and augmented with names found in the index records used in the project by Francis (2004). A table of equivalent birthplaces due to boundary changes after World War I was provided by Jensen (2004).

3.2 Establishing a Threshold Value for Scores and Estimating Error Rates

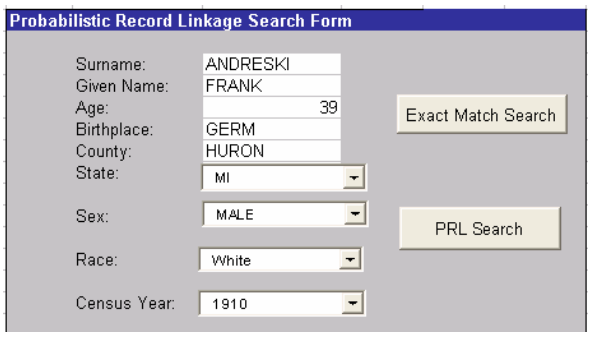
When searching for an ancestor in the census index, it is better to come up with a list of potential records than it is to miss the single record that represents your ancestor. For this reason we chose a single threshold for scores and sought to minimize the false negative linkage error rate ($\lambda = P(\text{not linked}|M)$). To establish a threshold, initially we used something similar to the SimRate approach of Winglee, Valliant and Scheuren (2005). Instead of simulating the distribution of scores, we enumerated the scores for all possible $5 \times 3^3 \times 2$ combinations of agreement status using the weights that we determined from the sample data in Section 2. The probability of obtaining each of these possible scores was calculated using both the estimated conditional probabilities, $\hat{m}_{v1}, \cdots, \hat{m}_{vc_v}$, and $\hat{u}_{v1}, \cdots, \hat{u}_{vc_v}$ assuming independence of fields. The cumulative distribution of scores for both the matched and unmatched records was examined to establish a threshold. The larger the threshold, the lower the percentage of false matches and the higher the chances of missing the record sought. Using a threshold of 2 resulted in about a 2% chance of missing the record sought, with a 0.3% chance of a false positive.

No weights were calculated from the data for surname and gender, since these fields were used to block the records. However, if a search program required an exact match on these two fields, it is estimated that at least 5.6% (i.e., 1,017/18,263 from Table 1) of the matches would be missed since some pairs that represent the same person do not match on the surname and gender fields. In order to reduce the chances of missing a matched pair, we created ad-hoc weights for these fields to help in discrimination. The ad hoc weight for gender was created by making the positive weight for matching genders and the negative weight for different genders the same proportion of the positive and negative weights for given name that Yamagata (2000) found in a separate record linkage study where gender was not used as a blocking factor. Several degrees of *close* were used for the ad hoc surname weights using a simplified form of the string comparator suggested by Jaro (1989). A weight of +3.5 was assigned if surnames matched exactly. If surnames were different, but 60% or more of the letters matched, a weight of $(PM \times 10.6) - 7.1$ was assigned, where *PM* is the proportion of matching letters. If less than 60% surname letters matched, a weight of -7.1 was assigned.

By adding weights for surname and gender, we were able to increase the threshold to 2.5 while decreasing the chance of both missing the record sought (<1.3%) and the chance of false positives. This was confirmed by examining the empirical CDF of scores for a random sample of the pairs clerically identified to be a match previously, and a random sample of pairs that did not represent the same individual.

4. Implementation of Optimal Weights in a Prototype Search Program

Using the data described in Section 2, we created a prototype search program that could search for a record using the Probabilistic Record Linkage technique and the weights and threshold we created, in addition to the "exact match" strategy used on some web sites. The input screen for a prototype search program offers the user two alternatives. To use the exact match, consider searching for an ancestor named Frank Andreski in the 1910 census records. If you knew that Frank was 39 years old in 1910, white, male, born in Germany, and living in Huron County, Michigan, you would input this information in the search form as shown below.



If you click on the Exact Match Search button, you will find one record, as you would using the search program on the HeritageQuest Web site. The full census record lists Frank as a farmer with a large family of 10 children.

If you wanted to search for Frank in the 1920 census, change the age in the search form to 49 and the census year to 1920, but the program retrieves no results when you click on the Exact Match Search button. However, clicking on the PRL Search button, the following possible matches would be retrieved.

Surname	GName	Age	Gender	Race	BPlace	St	County
ANDERESKI	FRANK	49	M	W	POLA	MI	HURON
ANDERSON	FRANK	46	M	W	MI	MI	OCEANA
ANDREWS	FRANK	53	M	W	IL	MI	HURON
SCHINIDER	FRANK	75	M	W	GERM	MI	HURON
SINDA	FRANK	48	M	W	MI	MI	HURON

The first entry matches the input form closest. The name is spelled differently (Andereski rather than Andreski) and the birthplace has changed from Germany to Poland. If you input this spelling of the name and birthplace into the advanced search form at HeritageQuest Online and follow the link, you will see from the image of the census record that this is the same person as Frank Andreski in 1910. The names and ages of his wife and first 10 children match.

Due to boundary realignments, many German immigrants changed their birthplace after World War I, and this, along with the name misspelling, causes the exact match search program to fail. The prototype Probabilistic Record Linkage search program acknowledges possible changes in birthplace with the look-up table, and the weight for the misspelled name is still +2.32.

HeritageQuest online service, which is available through public libraries, is in the process of adding a Probabilistic Record Linkage search program using these weights.

5. EM Algorithm and Future Work

Jaro (1989) describes the use of the EM algorithm to estimate the conditional probabilities $m_{v1},\cdots,m_{vc_v}$ and $u_{v1},\cdots,u_{vc_v}$, in the case where there is no training set of record pairs that represents the same people in different files. His application was for the binomial case where each field comparison was classified as "same" or "different." We investigated the use of the EM algorithm for the multinomial case where several levels of agreement exist for each field comparison. We introduced the new variable $\mathbf{g}_r = (g_{rm}, g_{ru})$ where, for each pair of records, $g_{rm} = 1$, if $r \in M$, and $g_{ru} = 1$ if $r \in U$. In the case where no training set is available the group membership of each pair of records is unknown and $\mathbf{g}_r$ is considered to be missing. We start with initial estimates of $m_{v1},\cdots,m_{vc_v}$, $u_{v1},\cdots,u_{vc_v}$ and $\hat{p}$ the proportion of record pairs in *M*. In the E step of the EM algorithm we use Bayes Rule to find

$$\hat{g}_{rm} = \frac{\hat{p}\prod_{j=1}^{v}\prod_{i=1}^{c_v} m_{ij}^{y_{rij}}}{\hat{p}\prod_{j=1}^{v}\prod_{i=1}^{c_v} m_{ij}^{y_{rij}} + (1-\hat{p})\prod_{j=1}^{v}\prod_{i=1}^{c_v} u_{ij}^{y_{rij}}}$$

and similarly for g_{ru}. In the M step ML estimates of the probabilities are:

$$\hat{m}_{vi} = \frac{\sum_{i=1}^{R}\hat{g}_{rm}\,y_{rvi}}{\sum_{i=1}^{R}\hat{g}_{rm}}\ ,\ \hat{u}_{vi} = \frac{\sum_{i=1}^{R}\hat{g}_{ru}\,y_{rvi}}{\sum_{i=1}^{R}\hat{g}_{ru}},$$

$$\hat{p} = \frac{\sum_{i=1}^{R}\hat{g}_{rm}}{R}.$$

Bauman(2006) programmed the EM algorithm in SAS proc iml and applied it to the smallest data set from Louisiana. He found that the estimates of some of the $m_{v1},\cdots,m_{vc_v}$ could be off by as much 0.14 for a file this size. The accuracy was affected by the number of record pairs in *M*, the number of outcome events in the comparison of each field, and the percent of records with missing values. Simulation studies indicated that the error in estimates could be reduced to less than 0.02 if the number of record pairs was increased to 500,000.

We plan to obtain more census index records from Heritage Quest. They have records from the same counties in Texas and Virginia for the 1900 through 1930 census, which will include over one million original records. Since there are too many records to do clerical matching to define a training set, we will use the EM algorithm on this data to develop a set of

weights that will be useful for searching 1900-1930 census index records. It would also be interesting to explore a hierarchical Bayesian method proposed by Larsen(2002, 2005) to estimate weights. This method allows parameters to vary by block, and it could possibly be used to find weights that could vary from region to region in order to produce a more accurate search program.

6. Conclusions

With the availability of websites where information can be obtained about one's ancestors, online genealogical searches are quite common. However, current online search programs based on an exact match algorithm will miss the record sought 50% of the time due to inconsistencies in spellings and dates recorded in the electronic databases. With a simple implementation of Probabilistic Record Linkage, we have been able to significantly improve the potential of online searches over exact-match search programs. Future work can hopefully expand this from 1910-1920 to 1900-1930 and searches of other types of records that can be found electronically online.

References

Bauman, J.G.,(2006) "Computation of Weights for Probabilistic Record Linkage using the EM Algorithm" Unpublished M.S. Project, Department of Statistics, Brigham Young University.

Crume, R. (2004) *Plugging into Your Past-How to find real family history records online,* Betterway Books, Cincinnati, OH

Francis, M.A. (2004) "Probabilistic Record Linkage of Census Data," Unpublished M.S. Project, Department of Statistics, Brigham Young University.

Ivan B. Fellegi, and Alan B. Sunter (1969) "A Theory for Record Linkage," *Journal of the American Statistical Association* 64: 1321-32.

Jensen, K. P. (2004) "Probabilistic Methodology for Record Linkage: Determining Robustness of Weights", Unpublished M.S. Project, Department of Statistics, Brigham Young University.

Jaro, M. A. (1989) "Advances in Record Linkage Methodology as Applied to the 1985 Census of Tampa Florida," *Journal of the American Statistical Association*, 84, 414-420.

Kemp,T. J., (2003) *Virtual Roots 2.0-A guide to genealogy and local history on the world wide web*, SR Books, Washington, DE

Larsen, Michael D. (2002) "Comments on Hierarchical Bayesian Record Linkage", *ASA Proceedings of the Survey Research Methods Section*, pp. 1995-2000.

Larsen, Michael D. (2005) "Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory", *ASA Proceedings of the Survey Research Methods Section*, pp. 3277-3284.

Lawson, J. (2006) "Find Them in The Census Records–using advanced record linkage techniques," *Everton's Genealogical Helper*, July-August 2006, 121-128.

Maritz. (2000, May 16). Recent Maritz poll shows explosion in popularity of genealogy. Public Relations Newswire. [Summary data available from Maritz Marketing Research] "http://www.genealogy.com /genealogy/ press-051600.html"

Yamagata (2001) "Probabilistic Methodology for Genealogical Record Linkage Increasing Classification Rates and Decreasing Unclassified Rates, Unpublished M.S. Project, Department of Statistics, Brigham Young University.

Winglee, M. Valliant, R and Scheuren, F. (2005) "A Case Study in Record Linkage" *Survey Methodology*, Vol. 31, No.1, pp. 3-11.