PROBABILISTIC METHODOLOGY FOR RECORD LINKAGE DETERMINING ROBUSTNESS OF WEIGHTS

By:
Krista P. Jensen, Dr. John S Lawson

Over time, the world population has developed a desire to research their ancestoral linage. Many resources have been identified to aid an individual in genealogical research. In the United States, one of the greatest resources for researching genealogy is census records. Census records allow a genealogical researcher to track individuals over time, broadening the scope of information one can acquire about an individual.

Dr. Halbert Dunn first presented the concept of *record linkage* in 1946 to describe the process, which joins two separate pieces of information for a particular individual or family [Dunn 1946]. Later, Fellegi and Sunter [1969] built upon Dunn's foundations by establishing a probabilistic mathematical approach to record linkage.

Probabilistic methods for record linkage have been developed to mimic the decision process of genealogists and researchers. An automated probabilistic approach allows the researcher to conduct many different types of searches within seconds. Following an automated search a list of record matches (links) as well as potential matches (links) with information necessary to further explore each potential record pair can be made. This enables a researcher to compile large numbers of records in a fraction of the time it would take to process manually.

Probabilistic methods have been applied to determine the feasibility of linking persons across multiple census years. With a set of known weights to use in the record linkage process, one would eliminate the need to examine a large number of records manually. This paper uses probabilistic methods to link census records from the 1910 and the 1920 census indices to illustrate the benefits of an automated record linkage approach.

CENSUS INDICES

Since before 1850, census records have provided information regarding one's demographic and personal information. Census indices contain a subset of the information found on a census page. In addition to omitting some information, census indices only include records for the head of household and individuals that differ in last name from the head of household [Szucs 2001]. Because of the limited information available in a census index, the defining demographics to be used in record linkage are likewise limited.

The subset of information found in a census index is as follows: surname, given name (sometimes a middle name or initial is present in the field for given name), age at the time of census, gender, race, country of origin, state of residence, county of residence, district of residence, and information about the census page the information is located.

When linking census records from any time period, it is important to account for discrepancies between censuses and failings inherent in the censuses. Because records from 1910 and 1920 have been used herein, issues relating to these census years will be presented.

In 1918, at the end of World War I many Eastern European boundaries were realigned, changing the "place of origin" for many immigrants in the United States. For instance, an individual listing their "place of origin" as Prussia in1910 would list Germany in 1920. Though the country of origin is typically stated, many instances arise where a region or city is given instead of a country, like Bavaria, a major region in Southern Germany.

The most prominent concern in using census records is the reliability of the reported age. Many individuals were secretive about their age or were unaware of their actual birth date.

When an individual did not know his or her exact age, it was rounded to the nearest decade by the enumerator. This resulted in discrepancies when tracking individuals over multiple census records. In one case, a woman only aged twelve years over a 30-year period [Szucs 2001]. Also, censuses were not always taken in the same month, as is the case for censuses taken in 1910 and 1920. For the 1910 census, individuals were instructed to give their age as of 1 April, 1910 and for 1920, they were instructed to give their age as of 1 January, 1920. This creates a one-year discrepancy in the number of years aged for individuals with birthdays between January and April.

METHODS

Record linkage presented by Fellegi and Sunter assumes two populations (data sets) are present and that some elements are common to both populations (data sets). When comparing two data sets, every record comparison is assigned one of three criteria: (1) Match (positive link), (2) Unmatched (positive non-link), (3) Possible link or Undetermined.

The set of true matches and true non-matches are denoted as *M* and *U* respectively. Comparing two data sets derives patterns of agreement between record pairs; these patterns are identified as the conditional probability P(*M*) and the unconditional probability P(*U*) (the observed pattern if a record pair is a match or the record pair is a non-match).

Methodology proposed by Dr. David White bases weights on a probability that two records being compared are a match given a certain event occurs, $P(M|E)$ [White 1997]. These events are (1) fields are the same or (2) fields are different. Therefore, two conditional probabilities are needed: the probability records are a match given a certain field is equivalent in both records, $P(M|S)$, and the probability records are a match given a certain field differs in the two records, $P(M|D)$.

Using a data set with known matches, $P(M|S)$ and $P(M|D)$ can be calculated. Conditional probabilities $P(S|M)$ and $P(D|M)$ are calculated by assessing all matches and counting the number of times fields are the same and different. The unconditional probability $P(M)$ is found by counting the number of matches. Unconditional probabilities $P(S)$ and $P(D)$ are estimated by taking a random sample of pairs and summing the times the fields are the same or different.

The set of known matches was identified by hand. It is known that manually linked records will have some degree of inherent error. However, these inaccuracies are only important if they substantially alter the calculated frequencies.

Weights were calculated for each field based on calculated conditional and unconditional probabilities. The application of weights was dependent upon the classification status of fields within a pair of records. If the entries for a particular field matched they received the weight:

$$w_k = \ln\left(\frac{P(Same \mid Match)}{P(Same_k)}\right)$$

A score or test statistic *W* for each record pair is simply the sum of the weights for each of the independent fields. Where W is

$$W = \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \ln\left(\frac{P(e_i \mid M)}{P(e_i)}\right).$$

Combing of similar records into groups (blocking) reduces the number of record comparisons to be made. Record comparisons are restricted to records within a given block decreasing the number of comparisons to be made.

In order to determine whether a record pair is considered a match, threshold values are used as the criteria in the classification of each record pair. The threshold values T_{μ} and T_{λ} are simply the weights $w(T_{\mu})$ and $w(T_{\lambda})$. These values are determined by a researcher to maximize the number of positively matched records and positively non-matched records. When determining the threshold values it is important to consider the two error rates. The first error rate is defined as false matches (non-matches that are classified as matches). The second error rate is defined as false non-matches (matches that are classified as non-matches).

RESULTS

Heritage Quest provided indexed census files from the 1910 and 1920 decennial censuses of five states: California, Connecticut, Illinois, Michigan and Louisiana.

California was split into two groups based on geographical region. One group contains data from Northern California while the other contains data from Southern California.
The set of known matches for the six groups of data were determined manually. The number of known matches for each data set ranged from 596 for Louisiana to 4,984 for Illinois.

To take into account the discrepancies in age, an algorithm was created to identify specific ranges of age differences and classify them as either a match (8-12 years aged), close (aged 7 or 13 years), or non-match (aged less than 7 years or aged more than 13 years). Border changes after World War I were researched as well as prominent cities and their national locations. A table listing this research was created to account for origination discrepancies in the field “place of origin”. First names were compared using the first three letters of a given name, the last three letters of the given name, and the first letter of a given name as well as the use of nicknames. This strategy was used to take into account variations of similar names.

An example of a record pair and the field classifications is found in Table 1:

Table 1: Record Comparison

Census Year	Surname	Given Name	Age	Gender	Race	Origin	State	County	District
1920-8780	DRECHSKER	OTTO C	48	M	W	SAXO	CT	TOLLAND	4-WD ROCKVILLE VERNON
1910-2334	DRECHSLER	OTTO	38	M	W	GERM	CT	TOLLAND	4-WD ROCKVILLE

The surname would be classified as different, thus eliminating it from comparison because of the blocking scheme. If these records were identified as a match, the score calculation would be as follows:

Given name - match, Age - match, Gender - match, Race - match, Origin - match, State match, County - match and District - match. Provides a score of

$$\mathbf{4.18 + 2.45 + 0.18 - 2.67 + 2.02 + 0.50 + 2.02 = 6.498}$$

This score identifies the record pair as a positive link (match) given the threshold value of either 1.806 or 2.504.

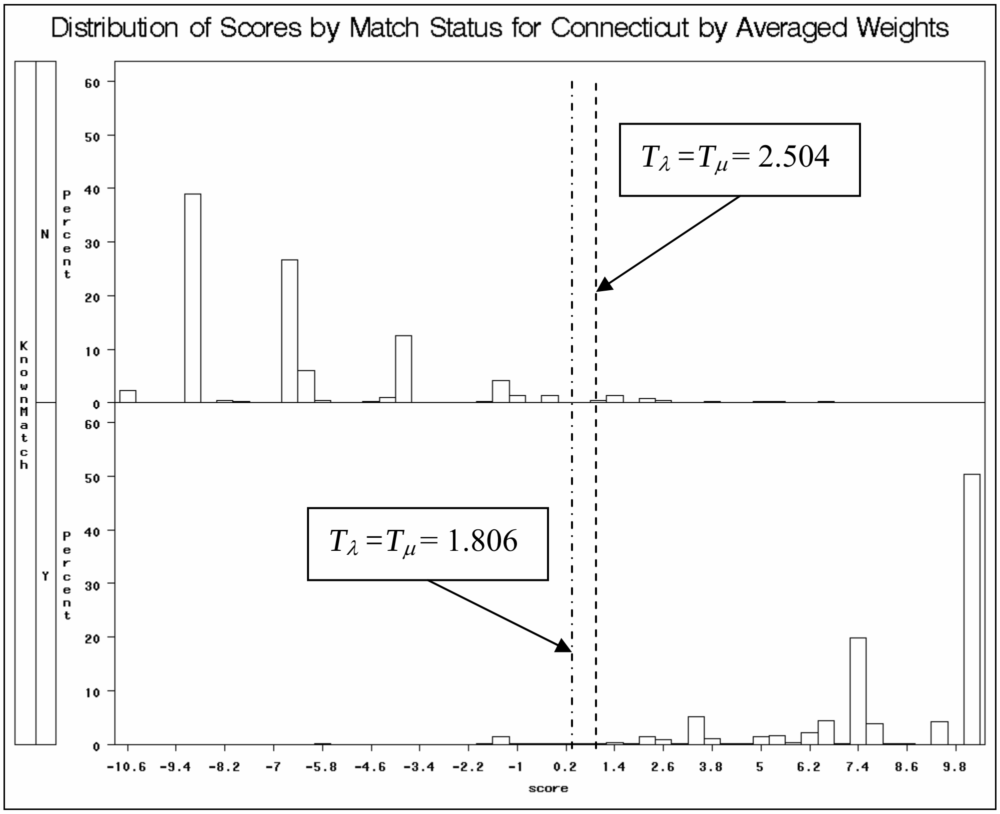
Weights were first calculated for the individual data sets and applied to the remaining sets. The results for three of the weight groups were promising but did not provide adequate results when matching records in every data set. None of the data sets used adequately represented the overall demographics of the United States. (For example, Louisiana has more individuals of French heritage, whereas a large Asian base is evident in both California data sets.) With strong fluctuations of ethnicities to specific geographical regions, it was theorized that an average of the weights would account for geographic variations in ethnicity. When tested, this theory was proven correct as evidenced by an error rate less than the standard level of 0.05.

Weights were calculated for each data set and then averaged. The listing of calculated averaged weights can be found in Table 2. After weights were calculated they were applied to each data set to determine their efficiency.

Table 2: Averaged Weights

Averaged: **Fields**	**Weight for "Same"**	**Weight for "Close"**	**Weight for "Different"**
Given Name	4.180092	-1.25993	-4.76084
First 3 letters of Given Name		3.3928	
First letter of Given Name		0.356995	
Last 3 letters of Given Name		-0.22506	
Age	2.455072	-0.10778	-2.63094
Race	0.183053	0.843567	-1.58802
Place or Origin	1.49957	-0.95751	-2.66818
Locale of Census	2.02468	1.521342	-1.35869
County	0.502542		-3.16472

Two threshold values were chosen for comparison, where $T_{\mu} = T_{\lambda}$. By choosing a threshold value where $T_{\mu} = T_{\lambda}$ there are no unclassified records. The threshold values chosen are $T_{\mu} = T_{\lambda}$=2.504 & 1.806. An illustration of weight distribution for the Connecticut data set can be found in Figure 1. It can be seen that good separation of matched and non-matched record pairs was achieved using the averaged weights.



$T_{\lambda} = T_{\mu} = 2.504$

$T_{\lambda} = T_{\mu} = 1.806$

Figure 1: Match Status Distribution of Connecticut Records using Averaged Weights

The error rates obtained using the averaged weights are all below the 0.05 level and have above 95% matching rates. A table listing the rates of matching relative to the calculated weights used in the record linkage process for both "Matched" records and "Non-Matched" records is given in Table 3. Table 4 provides the error rates associated with each data set.

Table 3: Percentage Linked Records

Linkage Percentages		Weights	
		Averaged Weights	Averaged Weights
Record Set	Threshold	Matches	Non-Matches
Louisiana	1.806	98.727	97.636
	2.504	98.02	98.648
Michigan	1.806	98.436	98.674
	2.504	97.599	99.088
Illinois	1.806	98.716	96.066
	2.504	98.154	97.169
Connecticut	1.806	97.505	97.446
	2.504	96.091	98.287
Southern California	1.806	97.805	98.25
	2.504	96.725	98.646
Northern California	1.806	98.919	98.688
	2.504	98.096	99.148

Table 4: Error Rates for λ & μ

Error Rates		Weights	
		Averaged Weights	Averaged Weights
Record Set	Threshold	λ	μ
Louisiana	1.806	0.0198	0.02359
	2.504	0.02687	0.01346
Michigan	1.806	0.01763	0.01316
	2.504	0.026	0.00902
Illinois	1.806	0.00522	0.01233
	2.504	0.01083	0.0141
Connecticut	1.806	0.01414	0.02542
	2.504	0.04075	0.01701
Southern California	1.806	0.03203	0.0169
	2.504	0.04282	0.01293
Northern California	1.806	0.01338	0.01302
	2.504	0.02162	0.00842

The averaged weights provide low error rates and high matching rates. They also allow a weight to be present for fields of interest present in one data set and not in the other (i.e. county or race). Misclassification of records is attributed to problems in identifying potential matches through given name. Approximately 92% of the misclassified matched records were record pairs that had different given names that could not be accounted for by any of the fields of choice. One of the most common causes of misclassification was a transposition of the first and middle names or initials. Others were obvious misspellings of the given name that could not be picked up by the algorithm used in this analysis.

In the mis-classification of non-matches, the major issues were record pairs that were similar in most fields but had different given names that had the same first letter. Because the given name wasn't classified as different, the negative weight for given name was not applied to the score of the record pair, giving the record pair a higher score and increasing the probability of being classified as a match instead of a non-match. The draw back of eliminating the first letter match occurs when one record reports only an initial and the other record gives a full name. Without a first letter match the records would receive a negative weight for "different" and not be selected as a match or a potential match.

When the averaged weights were applied to the six data sets available, error rates below the standard 0.05 level were obtained. The classification rates were all above 95% with the majority of errors being identified as failures inherent in the algorithm.

This paper shows that though weight sets cannot be interchanged effectively between regions (states), an average set of weights from a sample of regions can be used interchangeably with each other. It is presumed that this interchangeability holds for censuses across the nation for the same time period. It is theorized that using an average weight set from one time period (i.e. pre-WWII) to analyze census records in another (i.e. post WWII) would be less effective due to population fluctuations sensitive to politics, human migratory patterns and economic conditions particular to specific time periods.

It has been shown that it is possible to calculate a set of average weights from a sample of geographical areas that is valid for all record sets for different regions of a given time period. Analysis suggests this average weight set is only valid for a given time period due to sensitivity to historical factors influencing human migration. Considering an error rate of only 0.05, this record linkage tool will undoubtedly become useful to genealogists.

REFERENCES

Cerny, J. (1985), *Guide to Research, Case studies in American Genealogy*, Salt Lake City: Ancestry Incorporated

Columbia Gazetteer pf the World. Cohen, S.B., ed. New York: Columbia University Press, 1998. 3 vols.

Dunn, H.L. (1946), Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

Francis, M. (2003), Probabilistic Record Linkage of Census Data

Fellegi, I.P. Sunter, A.B. (1969), A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Kofler, M. (2000), *Definitive Guide to Excel VBA*, trans. David Kramer. New York: Springer-Verlag. Berkeley: Apress.

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

Newcombe, H.B., Kennedy, J.M., Axford, S.J., A.P. (1959), Automatic Linkage of Vital Records, Science, 130, 954-959.

Newcombe, H.B., (1988), *Handbook of Record Linkage*, New York: Oxford University Press.

Szucs, Loretto, Dennis, Wright, Matthew, (2001), *Finding Answers in U.S. Census Records*: Utah: MyFamily.com

Thorndale, W., and Dollarhide W. *Map Guide to the U.S. Federal Censuses*, 1790-1920. Baltimore: Genealogical Publishing Co., 1987.

U.S. Census Bureau (2002). *Measuring America: The Decennial Censuses From 1790 to 2000*. http://www.census.gov/prod/2002pubs/pol02-ma.pdf

White, D. (1997), A Review of the Statistics of Record Linkage for Genealogical Research, *Record Linkage Techniques-1997: Proceedings of an International Workshop and Exposition*, pp. 362- 373. http://www.fcsm.gov/working-papers/dwhite.pdf

Winkler, W.E. (1994), Advanced Methods for Record Linkage, *Proceedings of the* Section on Survey Research Methods, American Statistical Association, pp. 467-472.