Published in final edited form as:
Hist Methods. 2011 January 1; 44(1): 7–14. doi:10.1080/01615440.2010.517152.

New Methods of Census Record Linking

Ron Goeken, **Lap Huynh**, **Thomas Lenius**, and **Rebecca Vick**
Minnesota Population Center

Abstract

The Minnesota Population Center (MPC) has released linked datasets through its NAPP and IPUMS projects, making them readily accessible to researchers. Prior to the availability of complete count census microdata from the MPC, researchers applied various forms of record-linking software. This essay describes the techniques used in the MPC's linking program and briefly compares this technique with those used by other researchers. The key feature of the MPC linking method is the construction of cumulative name similarity scores, based on approximately 2.5 billion record comparisons; we also use support vector mechanics to classify potential links. This article explains modifications made for the final linked datasets and includes a discussion of the role of weighting variables when using linked data.

Keywords

census samples; complete count; record linkage; microdata; historical demography

The Minnesota Population Center has created a set of linked representative samples of individuals and family groups using U.S. Census microdata for the period 1850 through 1930, and we plan similar linked samples using data from Britain, Canada, Iceland, and Norway. The IPUMS nationally representative samples of the United States population from 1850 to the present have motivated much research on changing demographic and social behavior. A basic limitation of the existing samples, however, is that they are cross-sectional snapshots that do not allow observation of the same individuals in different census years. Linked data, in contrast, allow researchers to more directly and reliably examine topics like family formation and dissolution, social and geographic mobility, and the interrelationship of geographic and economic movement (Ruggles 2006).[1]

Record linkage has long been a basic tool of quantitative historical analysis. Traditionally, historical investigations focused on a single community, and records were linked by hand. This meant that persons who moved across the boundaries of the localities were lost, making the linked datasets non-representative of the broader population (Thernstrom 1964; Katz 1975; Knights 1991). In an effort to overcome this limitation, several studies used “soundex” name indexes to link individuals who had migrated (Ferrie 1996; Guest 1987; Steckel 1987). But these results were also mixed; the indexes are state-specific, so searching for migrants then required consulting each state index for a potential match. For every potential match located in the name indexes, the investigators had to manually consult microfilm of the census manuscripts to verify links. The resulting linked samples were comparatively small and expensive, and questions concerning representativeness remained.

Work on this project was completed under “Population Database for the United States,” National Institutes of Health, 5R01-HD039327.

[1]For more information on the 1880 complete-count database see http://www.nappdata.org/napp/. Information on the non-1880 samples and the linked samples can be found at http://usa.ipums.org/usa/sampdesc.shtml).

This paper describes the record-linkage procedures developed by the Minnesota Population Center to build better linked census samples than have previously been available.

The Minnesota Population Center Linkage Project

While all record linkage projects share a common goal of tracing individuals across time, decisions made about what information is used for linking and what constitutes an acceptable link vary across such projects. The MPC linkage project sought to satisfy the sometimes competing goals of creating links that were accurate and representative of the general population.

To maximize representativeness, we used a minimal set of linking variables. The source data consist of complete households, with information available for all co-resident household members. A record linkage algorithm that used information on co-resident household members would result in higher linkage rates and more accurate links. However, these benefits come at a cost—persons who remained with the same kin between census years would be overrrepresented in the linked data. Such bias in the linked data would yield erroneous conclusions about family transitions. Similarly, it would be inappropriate to consider place of residence or occupation when linking persons across censuses, since this would introduce bias with respect to migration or occupational mobility.[2]

To minimize selection bias, we restricted the linking variables to an individual's given name, surname, and birth year. For men and for women who do not marry between the censuses, these characteristics generally do not change over time, and therefore have the potential to yield a representative dataset.

Successful record linkage requires a mechanism for assessing name and age similarity. The ability to assess similarity can be enhanced by cleaning and standardizing the source data. We initially cleaned and edited the data prior to release as part of the IPUMS (Block and Star, 1995; Goeken, et al. 2003).[3] Age information, for example, is subject to a variety of consistency checks at the original data collection stage and later in IPUMS processing. The name fields, in contrast, receive little processing prior to IPUMS release.

We processed the surname field minimally. Non-alpha characters were removed, but there was no attempt to standardize or correct perceived misspellings. We did more processing of given names, by standardizing the first name strings that occurred at least 100 times. This name standardization process is described in detail in Vick (2010, in this issue).

We used Freely Extensible Biomedical Record Linkage (FEBRL) software to construct name and age similarity scores (Christien and Churches 2005). We extracted records from our databases based on race and birthplace, with separate files for males, females, and married couples. We then compared files from two censuses for persons of the same race, sex, and birthplace. For example, we compared white males born in Michigan in the 1870 data with white males born in Michigan in the 1880 data.[4] We further restricted comparisons to persons born within seven years of one another. For each comparison, we calculated

[2]There is insufficient information to correct for biases in the linkage process. It is relatively straightforward to construct weights that reflect age or occupational distributions. However, since we do not know the migration status for individual records, we have no way to construct weights that would correct for bias caused by using place of residence as a linking variable.

[3]See http://usa.ipums.org/usa/doc.shtml for a discussion of IPUMS logical edit procedures.

[4]Eligibility to be linked is dependent on being present in a given sample year and the 1880 complete-count database. For linked couples, the linkable population is also restricted by the requirement of being married and co-resident in both the sample year and 1880. For females, the linkable population consists of women who did not change t heir surname between sample year and 1880. Thus, for females we cannot link those who transitioned from single to married, nor those that remarry. There are no comparable restrictions on the male linkable population.

similarity scores; when a potential match exceeded minimum thresholds, the record pair was written to a results file.

After constructing similarity scores we evaluated the potential links.[5] To classify potential links as "true" or `false" we used a machine-learning tool known as SVM—Support Vector Machine (Christianni, Nello and Shawe-Taylor 2000; Steinwart and Christmann 2008; Pamarthy 2007). SVM construction depends on the existence of training data, which typically consists of a verified set of true and false links.[6] The SVM classifier analyzes the training data, plots them in a multidimensional space, and then constructs a boundary between the two classes of records that maximizes the distance from the hyperplane and the nearest data points in both of the classes (i.e., between the true and false links). The end result is a file consisting of potential links and the classifier-produced confidence score. Confidence scores are interpreted dichotomously; a positive score is considered a "true" link and a negative score is a "false" link.

At the classifier stage each potential link is evaluated independently, which often results in numerous potential links from the 1880 complete census to a given record from an IPUMS sample. We considered any case with multiple potential links to be ambiguous and reject them. Table 1 shows the confidence scores for potential links to John Bradley, a 25-year-old white male born in South Carolina from the 1870 data. Of the 43 potential links, only the top four receive positive confidence scores. Although the potential link with the highest confidence score is an exact match, the other three also have a high degree of similarity. If we had to choose, the exact match is probably the correct link. However, our analysis of such cases suggests that the probability that the top link is the correct link is significantly under 95 percent, and using these types of links would introduce an unacceptable error rate.

Preliminary Linked Files and Subsequent Linkage Process Modifications

In fall 2008, we released preliminary versions of the linked samples that were created using the procedures outlined above. We anticipated making improvements for the final release, and part of that process was comparing our links to a set of links for 1870 and 1900 produced by Pleiades Software Development.[7] Pleiades produces record linkage software designed for genealogical research and has been involved in numerous record linkage projects over the past 20 years. Their linkage process is based on an additive point system that assesses similarity for individual records. In contrast to our approach, Pleiades' system uses household and residential information (which we excluded as sources of potential bias, as explained above).

Among the native-born white males present in both linked samples, the MPC linked samples agreed with the Pleiades links 98.8 percent of the time—a reassuring result. Only 44 percent of our preliminary links, however, were in the Pleiades linked set. Visual examination of the household data for the records that were only present in our linked data disclosed few that seemed ambiguous or likely errors; the overwhelming majority appeared to be accurate links.

[5]We also construct variables based on individual-level characteristics at this point. Although most of this work involves married couples--e.g., whether both spouses had ages ending in zero, or whether the husband is older than the wife in one year, but younger in the other year—we also construct phonetic codes for all last names.

[6]In practice, however, few linkage projects have verified training data. For our project, we selected a random sample of potential links, and had a group of MPC data entry operators code each potential link as a "yes" or "no" based on a visual examination of names and ages of potential links (with yes indicating that it was in their opinion a true link). If a majority had the potential link as a "yes", then it was coded as a "yes" in the training data (with the remainder coded as "no").

[7]See: http://www.pleiades-software.com/

Why were so many of the MPC-only links missing from the Pleiades linked dataset? We discovered that the MPC-only links were likely to be young individuals who were enumerated as a child in both 1870 and 1880. A review of these links disclosed that parents in these households often had imprecise or conflicting information. Examples would be households that had transitioned from couple-headed in 1870 to a single parent head in 1880, or that had imprecise parental name, age, or birthplace information. The comparison of our links to the Pleiades set thus highlighted another disadvantage of using household information to conduct record linkage: such information increases linkage rates only in cases where household information is consistent over time.[8]

Of more direct relevance to the MPC linkage project was the evidence of a high degree of imprecision in nineteenth century census data. Our data consist of primary links (of individuals), and we then proceed to link other household members after the primary links are established. Under our preliminary linking rules, primary links generally had to be within two years of expected age. But a comparison of links for the 1870–1880 male sample showed that approximately 10 percent of the household links had an 1880 age that was more than 2 years different than their expected age.[9]

The Pleiades comparison and the evidence regarding high levels of age imprecision led us to reexamine a fundamental part of our linking process. The issue was that our linkage process was generally accurate, but was also highly dependent on the precision of the data. More specifically, if the individuals we attempted to link were accurately enumerated in both census years, we would either make the link or, in the case of multiple potential links, reject the link as ambiguous. But if the correct link was unidentifiable—because of mortality, under-enumeration, or mis-enumeration of linking variable information—we would make an incorrect link if there were another person with similar characteristics in the 1880 complete-count data.

We dealt with this problem by developing formal measures for summarizing and taking into account the commonness of names. A fairly standard approach in record linkage projects is to construct frequency tables for names, which more or less assess the probability of a correct link. For example, based on frequency tables, record linkers would be more confident in linking someone with the name Roland Marsupial than someone named John Smith. But given our minimalist approach to name cleaning and standardization, we ran into a problem with minor typos and misspellings which would show up as low frequency names, even though many of these names have high similarity to high frequency names (and would show up as potential links to records with high frequency names). Our solution was to construct name similarity scores based on the following: for a given sample record, we determined the proportion of records (by race, birthplace, and sex) in the 1880 complete-count data with a Jaro-Winkler similarity score greater than 0.9. The choice of this threshold is somewhat arbitrary, but, based on the preliminary linked data, we rarely linked records that did not exceed this threshold.

Just as it is more difficult to link individuals with common names, it is more difficult to link individuals born into states with large populations. With this in mind, we also constructed a

[8]This problem extends beyond overall linkage rates to issues relating to accuracy. For example, a “true” link can be rejected because of presence of internal household disagreement, while a “false” linking can be accepted due to the absence of such household disagreement.

[9]It is difficult to say whether this an over or underestimate of age precision. The presence of incorrect links could inflate this measure, because if the primary link is incorrect, then we would also assume any co-resident household links would also be incorrect (and would thus have a high likelihood of age imprecision). However, the establishment of a primary link typically means that at least one person in a given household had enough age precision to be linked, and we assume that this would be correlated with age precision for other household members.

density of birth measure, which is the proportion of 1880 records for specific birthplaces, by race and sex. Our expectation was that we would rarely (if ever) link records with common names from the larger states of birth (like New York and Pennsylvania), but we would be able to link relatively common names taken from the smaller states of birth (such as Delaware).

In addition to name commonness and birthplace density measures, we constructed new training data based on the Pleiades linked data. We also expanded our linkage variable set to take into account middle name (or initial) agreement, as well as parental birthplace consistency, which is only available beginning with the 1880 census. The new linked results were characterized as having high levels of similarity along with significantly higher linkage rates. But we suspected (largely based on migration differentials for different classes of linked records) that we were adding an unacceptable number of incorrect links. The basic problem was that our new classifiers were “tight.” Exact and near matches continued to be classified as true links, but we were no longer considering less precise matches as true links. As a result, some cases that should have been rejected as ambiguous were being considered valid links. We solved this problem by developing two classifiers—one of which was tight and the other loose—and we conservatively defined true links as records that had one and only one positive link in both models.

Differential Linking Results by Race, Nativity, and Birthplace

African-Americans and foreign-born whites had considerably lower linkage rates than did native-born whites. This is partly because of differentials in misreporting of age. The African-American population had high age misreporting (Elo and Preston 1994; Coale and Rives 1973), and age-heaping evidence suggests that immigrant groups also had higher levels of age misreporting than did native-born whites.

A second explanation for differential linkage rates is different degrees of name homogeneity across population groups. As noted, common names produce multiple ambiguous links and we excluded all such ambiguous links to avoid incorporating false links into the dataset. If certain racial and ethnic groups tend to be clustered into common name categories, then the final linkage rates for those groups will be correspondingly lower.

We classified 18 percent of the foreign-born in our highest name commonness category, compared with 10 percent of African-Americans and only 7 percent of native-born whites. Name commonness also varies by country of birth. For example, only 2 percent of white males born in France had names falling into the most common category, versus 63 percent of white males born in Wales. Other immigrant groups with high name homogeneity were males born in Ireland, Scotland, Norway, and Sweden, with between 32 and 40 percent of these nationalities having most common names. Linkage rates are powerfully inversely associated with name commonness. For example, among native-born whites the linkage rate between 1870 and 1880 was 19 percent for those with the least common names, and just 0.6 percent for those with the most common names.

A third factor affecting linkage rates is birthplace (and the population of an individual's state of birth). For persons with moderately common names, the population of the birth state is inversely related to the linkage rate. That is, persons from small states are more often linked than those from large states simply because there is less chance that the link will be ambiguous.

We always assumed that we would deal with linkage differentials by constructing weights for the linked records, so that, all else being equal, linked individuals born in Delaware would have a lower weight than those born in New York. We did not anticipate correcting

biases resulting from the different linking rates for more versus less common names, however. That Roland Marsupial is more likely to be linked than John Smith would matter to researchers only if name commonness were systematically associated with other important characteristics, such as socioeconomic status. Fortunately, name commonness does not appear to be significant to other characteristics.

Table 2 gives the mean occupational scores for 1870 males by race/nativity and name commonness. Occupational score is an index of the earning power of each occupational title (Ruggles, et al. 2010). The first column of results (under mean OCCSCORE) excludes non-occupational responses to the occupation question, and the second column excludes both non-occupational responses and farmers.[10] There are, as would be expected, substantial differences in mean occupational scores between the race/ethnic groups, with African-American males substantially disadvantaged. But within race/ethnic categories, occupational scores are relatively flat across name categories. When non-occupational responses are excluded, the mean occupational score for native-born white males is 18.7 for those with the least common names and 18.5 for those with the most common.

Accuracy of Links in the Final Release

Thus far, this paper has addressed how links were made and how frequently links were made. Another important question is the accuracy of the links in the final samples released by the Minnesota Population Center.

Table 3 provides illustrative examples of three households from our 1870–1880 male linked sample. For both 1870 and 1880, the given name, surname, and relationship to head are provided, with linked individuals shown on the same line. “Linktype” indicates whether the record is a primary (individual) link or household link (made to persons co-residing with the primary linked individual).

In the first household shown, the primary link is the third individual (Alma). We linked the household members because of the high degree of name and age similarity. In the second household, the primary link is “Eddie Cimmerman” in 1870 and “Edward Zimmerman” in 1880. Although three members of the 1870 household are not present in 1880, we see high similarity among the other household members, despite the different surname spelling. The third household is an example of a primary link with a relatively rare given name (Duett). This contrasts with the household head's given name information; we would have difficulty linking two records enumerated as “L” and “Lathrop” in different census years in the absence of a rare given name. Once we have established the primary link (for Duett Manning), however, we will link the household head and other co-resident family members in the household linking process.

All primary links in Table 3 appear accurate despite some imprecision in the household links. In addition, all of these households remained in the same state and county in both census years, which increases our confidence that they were linked correctly. While it is difficult to estimate the number of incorrect links with accuracy, we used indirect measures to assess the general accuracy or consistency of the linking procedures. For example, in a comparison of married individuals in the linked males (or females) files with records in the couples linked files, we would expect to find comparable characteristics. To the extent that characteristics are not comparable, all else being equal it would reflect the higher accuracy of the couples links because of the addition of three extra variables to the couples linking

[10]OCCSCORE is an IPUMS constructed variable that assigns occupational income scores to specific occupations. See http://usa.ipums.org/

algorithm (given name, age, and birthplace of the spouse). For native-born white males in the 1870 couples data, we found 22.3 percent residing in a different state or different county in 1880. The corresponding migration rate for native-born white couples in the male and female only linked files were 23.0 and 23.9 percent, respectively.

We also identified male records that are linked in both the male and couples file to determine if they are linked to the same household in the 1880 data. The results are remarkably consistent. Of the 3,609 males in both the male and couples linked files, we link to different households only eight times.

As a final check on the accuracy of the links, we identified sets of brothers in the 1870 male sample that were enumerated as sons living with both parents in both years. Altogether we have 1,723 native-born whites in the 1870 male sample that satisfy this requirement. We would expect the specific sets of brothers to end up in the same household in 1880. This failed to occur in only 2.0 percent of the sets. The results of this analysis serve as an indirect error estimate for this group of links. Although some of the consistently linked sets of brothers could be errors, it would be rare to find inaccurate links among consistently linked records.

The estimate for linked sets of native-born white brothers, along with high levels of consistency for males linked in both the male and couples file, are important indicators of the accuracy of the MPC linking protocols. Taken together, the brothers and the married males linked in the couples file make up over 25 percent of all 1870–1880 male links. Error rates for foreign-born whites and African-American linked populations may be higher, but error rates for the native born linked populations are significantly below 5 percent.

Weighting of Samples

Table 4 shows the number of linked records for MPC linked samples by nativity/race and, for females, by nativity/race and marital status categories.[11] The number of records is low in some of the linked samples. Accordingly, we apply weights to compensate for under- and over-representation of population subgroups.

Population subgroups at particular risk for under-estimation include African Americans, immigrants, and boarders/lodgers. The imprecise age reporting and name homogeneity for African-Americans and the foreign born, along with the smaller absolute size of these population subgroups, account for the small number of links shown. However, this underrepresentation could also be influenced by respondent bias. Since enumerators went household to household and, presumably, spoke with household heads or the spouses of heads, it is unlikely that there was any direct communication with unrelated individuals (e.g., boarders, lodgers, and employees). This group—which was approximately 10 percent of the adult population in the nineteenth century—is underrepresented in our linked samples because of imprecise name, age, and possibly birthplace information. This bias is also reflected in the linkage rates for variables associated with the unrelated population (e.g., younger adults and residence in larger cities).

The MPC weighting strategy is to weight by population characteristics in the terminal year. Thus, for linked data from the 1850 to 1870 sample years, we weighted by 1880 characteristics. For the post-1880 sample years, we weighted by the sample year characteristics. Weights were based on an estimate of the “linkable” population. Using the 1870–1880 samples as an example, the linkable population for native-born groups is anyone

[11]The numbers given in Table 4 are for primary links. Weights are constructed only for these links.

who was 10 years or older in the 1880 census. Since we do not have year of immigration information before 1900, we had to look at the foreign-born in the 1870 census and apply life tables to estimate how many of these individuals would still have been alive in 1880. The difference between this estimate and the actual total in 1880 would be due to immigration between 1870 and 1880.

After identifying the appropriate estimates for linkable subgroups, we constructed an initial weight—which is the inverse of the groups linkage rate—that was used to inflate the linked sample to the actual (or estimated) population totals for all subgroups in the terminal year census. We then calculated the specific weights for a number of weighting variables. Again using 1870–1880 males as an example, we first estimated the relationship to head weight, which was calculated as the proportion of the linkable population by relationship-to-head categories divided by the proportions for the linked sample. After this we used the first weight to weight the linked records, and then calculated proportions for the next weighting variable (in this case individual birthplaces). We repeated this process for 5-year age groups, size of place and occupational categories, with a modification of each specific record's weight occurring with each iteration.

This process worked fairly well where we had enough records. However, the low subgroup case counts shown in Table 4 resulted in a large range of weight values, with some linked records representing a relatively small or large number of records in the original data. We addressed this problem by imposing a minimum and maximum on the weight for all subgroups. The minimum is one-fifth of the average weight for the subgroup, with the maximum capped at four times the average subgroup weigh. This had little effect on some larger groups; for example, less than 1 percent of native-born male links for 1870–1880 are affected by the minimum/maximum rule. However, almost 10 percent of the foreign-born records are either below or above the initial minimum or maximum.

Individual researchers must decide whether the constructed weights are appropriate for their specific study. Researchers should also take note of the small number of records for some subgroups listed in Table 4 and use caution when including these sub-groups in their studies.

Future Record Linkage Plans

We intend to alleviate the research problems imposed by small subgroup sample sizes by linking 5 percent samples for 1900 and 1930 to the 1880 complete-count database. This will increase the size of all subgroups listed in Table 4 by a factor of five. We also plan to expand the nineteenth-century-linked samples by a factor of 100. We are currently working on a complete-count database for the 1850 U. S. census. When this work is complete, we anticipate a subsequent project to use the 1850 complete-count data along with complete-count data for 1860, 1870, and 1900. This will greatly expand our current linked samples and also will allow us to create true longitudinal data. Linking individuals and their households across five different censuses—from 1850 to 1900—will transform our ability to understand nineteenth century population dynamics. Researchers do not need to wait, however. The linked samples described in this article are available now from the Minnesota Population Center.

References

Block WC, Star DL. Data entry and verification in the 1850, 1880 and 1920 Public Use Microdata Samples. Historical Methods. 1995; 28:63–65.

Christen, P.; Churches, T. Febrl - Freely extensible biomedical record linkage. (Manual, release 0.3), 0.3 edition. 2005.

Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press; London: 2000.

Coale AJ, Norfleet WR Jr. A statistical reconstruction of the black population of the United States 1880–1970: Estimates of true numbers by age and sex, birth rates, and total fertility. Population Index. 1972; 39:3–36.

Elo IT, Preston SH. Estimating African-American mortality from inaccurate data. Demogmphy. 1994; 31:427–58.

Ferrie JP. A new sample of males linked from the Public-Use-Microdata-Sample of the 1850 US Federal Census of Population to the 1860 US Federal Census manuscript schedules. Historical Methods. 1996; 29:141–156.

Goeken R, Nguyen C, Ruggles S, Sargent W. The 1880 United States population database. Historical Methods. 2003; 36:27–34.

Guest A. Notes from the National Panel Study: Linkage and migration in the late nineteenth century. Historical Methods. 1987; 20:63–77.

Katz, MB. The people of Hamilton, Canada west: Family and class in a mid-nineteenth-century city. Harvard University Press; Cambridge: 1975.

Knights, PR. Yankee destinies: The lives of ordinary nineteenth-century Bostonians. University of North Carolina Press; Chapel Hill: 1991.

Pamarthy, K. A machine learning framework for record linkage in census data. MS Report, University of Minnesota; 2007.

Ruggles S. Linking historical censuses: A new approach. History and Computing. 2006; 14:213–224.

Steckel R. Household migration and rural settlement in the United States, 1850–1860. Explorations in Economic History. 1987; 26:190–218.

Steinwart, I.; Christmann, A. Support vector machines. Springer-Verlag; New York: 2008.

Thernstrom, SA. Poverty and progress; social mobility in a nineteenth century city. Harvard University Press; Cambridge: 1964.

Table 1

Potential links and Confidence Scores for John Bradley

name1_70	namelast_70	name1_80	namelast_80	age70	age80	CONFIDENCE
john	bradley	john	bradley	25	35	1.163721561
john	bradley	john	bradly	25	34	0.999793589
john	bradley	john	bradley	25	37	0.999444664
john	bradley	john	bradley	25	38	0.879444664
john	bradley	h	bradley	25	35	−0.994843602
john	bradley	j	bailey	25	35	−0.995201766
john	bradley	john	shandley	25	34	−0.999585986
john	bradley	john	bryan	25	35	−0.999669075
john	bradley	john	ragsdalle	25	35	−1.000102878
john	bradley	john	bryante	25	35	−1.000563741
john	bradley	john	bail	25	35	−1.001999259
john	bradley	john	darby	25	36	−1.003973365
john	bradley	john	nalley	25	35	−1.010393977
john	bradley	john	ashley	25	35	−1.010393977
john	bradley	john	rarden	25	35	−1.010393977
john	bradley	john	ashley	25	35	−1.010393977
john	bradley	john	trader	25	35	−1.010393977
john	bradley	john	bryce	25	34	−1.011851192
john	bradley	john	ready	25	36	−1.019752145
john	bradley	john	beasley	25	35	−1.023576736
john	bradley	josiah	bramlet	25	35	−1.025904298
john	bradley	john	bayler	25	33	−1.027504802
john	bradley	john	blake	25	35	−1.028183818
john	bradley	john	boyer	25	35	−1.028183818
john	bradley	john	berry	25	35	−1.028183818
john	bradley	john	brownlee	25	36	−1.037933946
john	bradley	john	branch	25	34	−1.045258641

name1_70	namelast_70	name1_80	namelast_80	age70	age80	CONFIDENCE
john	bradley	john	clardy	25	34	−1.047429204

Table 2

Mean Occupational Score by Nativity/Race and Name Commonness Scores

	Excludes non-occupational responses			Excludes Farmers and non-occupational responses	
Nativity/Race	**Name category**	**Mean OCCSCORE**	**N**	**Mean OCCSCORE**	**N**
Native-born white	1 (least common)	18.71	19505	21.47	12285
	2	19.32	18200	22.26	11734
	3	19.40	7603	22.16	5033
	4	19.09	7257	21.61	4849
	5	19.28	7201	21.95	4778
	6	19.32	7224	21.91	4856
	7 (most common)	18.51	5667	20.91	3698
	Total	19.07	72657	21.80	47233
Foreign-born white	1 (least common)	22.11	5098	23.83	4208
	2	21.97	5329	23.96	4266
	3	21.99	2465	23.97	1976
	4	21.82	2378	23.47	1964
	5	22.14	2694	24.02	2189
	6	21.59	2814	23.32	2292
	7 (most common)	21.25	4739	22.70	3949
	Total	21.83	25517	23.59	20844
African-American	1 (least common)	12.43	7452	12.17	6368
	2	12.41	6299	12.14	5389
	3	12.65	2697	12.44	2324
	4	12.68	2576	12.47	2226
	5	12.85	2654	12.68	2317
	6	13.08	2769	12.94	2404
	7 (most common)	13.56	3527	13.51	3137
	Total	12.72	27974	12.51	24165

Table 3

Selected Linked Households, 1870–1880 Male Linked Sample

LINKTYPE	LAST70	FIRST70	LAST80	FIRST80	RELATE70	RELATE80	AGE70	AGE80
household	WHITE	JAMES D	WHITE	JAMES G.	Head	Head	50	60
household	WHITE	MARY	WHITE	MARY E.	Spouse	Spouse	31	41
primary	WHITE	ALVA	WHITE	ALVA D.	Son	Son	9	19
household	WHITE	EVA	WHITE	EVA	Daughter	Daughter	2	12
not linked			WHITE	JAMES J.		Son		22
household	CIMMERMAN	JOSEPH	ZIMMERMAN	JOSEPH	Head	Head	43	53
household	CIMMERMAN	CAROLINE	ZIMMERMAN	CAROLINE	Spouse	Spouse	43	53
not linked	CIMMERMAN	JOSEPH			Son		20	
not linked	CIMMERMAN	JOHN			Son		15	
not linked	CIMMERMAN	CAROLINE			Daughter		13	
primary	CIMMERMAN	EDDIE	ZIMMERMAN	EDWARD	Son	Son	10	20
household	CIMMERMAN	EMMA	ZIMMERMAN	EMMA	Daughter	Daughter	7	17
household	CIMMERMAN	LAURA	ZIMMERMAN	LAURA	Daughter	Daughter	4	14
household	MANNING	L	MANNING	LATHROP	Head	Head	58	68
household	MANNING	? ACENITH	MANNING	ASENATH	Spouse	Spouse	57	66
primary	MANNING	DUETT	MANNING	DUETT	Son	Son	16	26
not linked	WILSON	AGUSTUS			Unrelated		69	
not linked	WILSON	ELIZA			Unrelated		66	

Table 4

Number of Linked Records, by Sample Year and Linked Population Subgroup

MALE	nat-b white	for-b white	af-am
1850	7,013	299	82
1860	10,426	634	235
1870	17,725	879	2,180
1900	18,596	1,515	1,334
1910	14,855	995	791
1920	10,050	511	504
1930	9,018	336	352

FEMALE			
Native-born white	**married**	**single**	**formerly**
1850	1,077	468	798
1860	2,253	1,495	843
1870	4,254	6,700	1,134
1900	3,274	4,241	1,162
1910	1,891	1,407	1,124
1920	793	849	894
1930	221	700	545

Foreign-born white	married	single	formerly
1850	63	2	62
1860	215	31	79
1870	671	153	184
1900	554	65	252
1910	317	40	231
1920	111	17	145
1930	26	16	71

African-American	married	single	formerly
1850	8	9	12
1860	34	34	34
1870	432	899	134
1900	164	210	88
1910	73	40	51
1920	20	20	34
1930	2	7	29

COUPLES	nat-b white	for-b white	af-am
1850	2,135	227	6
1860	4,538	932	19
1870	8,862	2,267	407
1900	7,745	2,132	180

COUPLES	nat-b white	for-b white	af-am
1910	4,650	1,101	102
1920	2,107	416	26
1930	612	105	7