**Record linkage at the Minnesota Population Center**

Ron Goeken, Lap Huynh, Tom Lenius, and Rebecca Vick (Minnesota Population Center)

In 2003 the Minnesota Population Center was awarded a grant to create an IPUMS-compatible, 10-percent sample of the complete-count database of the 1880 United States population. The grant also proposed creating a series of independent linked samples consisting of married couples, males, and females.[[1]](#endnote-1) Each linked sample would use the complete-count database of the 1880 U.S. census and a sample of the United States population for the non-1880 census year. For example, the 1870 – 1880 linked samples would use the 1880 complete-count data and a 1-percent sample of the 1870 population.[[2]](#endnote-2)

The existence of nationally representative samples of the United States population have been very useful in motivating research on basic demographic and social behavior. However, a weakness of the existing samples is their cross-sectional nature; each of the IPUMS samples are independent (and contain very few common records). Linked data, in contrast, would allow researchers to more directly and reliably examine topics like family formation and dissolution, social and geographic mobility, the interrelationship of geographic and economic movement, and trends and differentials in social mobility.

Researchers have been linking historical census records for some time. A basic problem with the earliest attempts, which focused on specific localities and basically utilized hand-searching for links, was the inability to link individuals who moved.[[3]](#endnote-3) More recent linkage studies used “soundex” name indexes to facilitate the linking of individuals who had migrated.[[4]](#endnote-4) But these results were also mixed; soundex indexes existed for specific states, and searching for migrants then required consulting each state index for a potential match. The absence of machine readable complete-count data also meant that researchers had to consult microfilm (or more recently digitized image files) of the census manuscripts to locate potential links.

Despite these logistical difficulties, linking historical census data continued to attract interest because of the perceived benefits relating to basic social and demographic research. In addition, the obstacles to automating much of the process began to lessen with the availability of the 1880 complete-count database. Another motivating factor would be development in the field of computer science; much effort has been devoted to developing tools and algorithms designed to extract meaning from large bodies of data, and many of these innovations have applicability to record linkage projects.[[5]](#endnote-5)

***The Minnesota Population Center Linkage Project***

A basic advantage of automated record linkage over manual procedures is the ability to process potential links more efficiently. However, despite increases in computation speed, automated methods typically have to establish limits on the number of record comparisons—in our case, comparing every male in our 1870 one-percent sample to every male in the 1880 complete-count data would result in approximately 5 trillion comparisons. To minimize processing time, we limited comparisons to records that share the same sex, race and birthplace in their respective census years. Using the 1870 – 1880 male sample as an example, we only compared white males, born in Michigan in the 1870 data to white males, born in Michigan in the 1880 data. In addition to blocking the data by sex, race and birthplace, we also used a sliding age window to further restrict the number of record comparisons.[[6]](#endnote-6)

Given that there were individuals with incorrectly enumerated or transcribed information in the data, we would lose some potential links because of this decision. But another factor behind our decision to limit potential links to individuals with consistent race and birthplace information was the difficulty in determining whether a potential link was accurate if this information did not agree. A typical record linkage project might be willing to overlook race or birthplace inconsistency if other information was consistent and overwhelmingly indicated that the potential link was in fact a true link. For example, Norman Whitfield, a 27-year-old white male born in Ohio in the 1870 data could be the same person as Norman Whitfield, a 37-year-old white male born in Michigan in the 1870, especially if both individuals lived in the same state and county, and also if both individuals had a wife named Lavinia and children named Jeremiah and Emma.

However, in contrast to many data mining projects, we were more concerned with the accuracy and representativeness of our links, as opposed to maximizing our linkage rate. The data we use consists of complete households, with information available for all co-resident household members. A record linkage algorithm that takes into account the presence (or absence) of co-resident household members in two specific censuses would result in higher linkage rates. However, this also comes at a cost in that individuals living without kin become more difficult to link and would be underrepresented in the resulting data. Place of residence is another census variable that would be useful in the linking process. All things being equal, potential links residing in the same locality in successive censuses would be more likely to be accurate than potential links residing in different localities. But this would also result in migrants being underrepresented in the linked samples.[[7]](#endnote-7) Given our concerns regarding bias, mainly because we anticipated that a primary use of the linked data would be to examine topics like migration and family formation and dissolution, we restricted the linkage variables to an individual’s given name, surname and age.[[8]](#endnote-8)

 The decision to use a limited set of linkage variables meant that we needed a strategy for dealing with duplicates—i.e., individuals with identical names and ages within specific combinations of race and birthplace—in the 1880 complete-count data. The original grant proposed identifying and excluding duplicate records from the linking process. Table 1 gives the distribution of white males with the name John Smith, between the ages of 20 and 50, born in selected states from the 1880 complete-count database. For New York and Ohio, there are duplicates at all selected ages. For the other states there are a number of ages where we do not find duplicates. In Maryland, for example, we have only one John Smith at ages 31 and 44. In the 1870 1-percent sample we have only one white John Smith born in Maryland in this expected age range (1870 age plus 10); given that this John Smith was 20 years old in 1870, we would expect his age to be 30 in 1880. Eliminating the duplicate John Smiths in Maryland would result in a link between the 1870 John Smith (expected age of 30) and the only non-duplicate John Smith in 1880 (age of 31) if we were willing to tolerate an expected age difference of one year. This could be the correct link, but, depending on the age precision in the data, it is also plausible that the correct link could be any of the John Smiths that were 30 years old in the 1880 data, or any of the John Smiths that were 29 years old in the 1880 data.

Because we expected a fair amount of imprecision in the data, a process that eliminated duplicate records was rejected. Instead we would compare all records within race and birthplace blocks, and if we ultimately came up with more than one plausible link (from the 1880 data) for a given sample record, we would reject all of the potential links. What this ultimately meant was that the linked samples would largely consist of “unique” individuals; i.e., records from larger place of birth states (or countries) with fairly unique combinations of names and age, or records from smaller states of birth (where we find fewer duplicates). Given that we were primarily concerned with accuracy rather than maximizing linkage rates, this proved to be a viable strategy.

**Generating Similarity Scores and Classification of Potential Links**

Successful record linkage requires a mechanism for assessing name and age similarity. Exact matches are unambiguous but, as noted above, we anticipated accepting exact matches as true links only if we found no other potential links characterized as near matches. We also had to evaluate the similarity of respective name strings and ages in the absence of an exact match.

The ability to assess similarity can be enhanced by cleaning and standardizing the source data. The sample and complete-count data has been through a variety of cleaning and logical edits prior to release as part of the IPUMS.[[9]](#endnote-9) Age information, for example, is subject to a variety of consistency checks at the original data collection stage and later in IPUMS processing. Thus we felt no need to further process age information prior to linkage. The name fields, in contrast, receive little processing prior to IPUMS release. IPUMS data have separate fields for given and last name. While the last name field consistently contains a single string, the given name field can contain given and middle name, given name and middle initial, or even a lone first initial. Some enumerators also used abbreviations for common given names, which were transcribed verbatim in the data collection process.

We ultimately decided to do a minimal amount of processing on the surname field. We removed non-alpha characters, but did not attempt to standardize or correct perceived misspellings. We generally took the same approach with the given names in that we were not overly concerned with misspellings. For example, we felt that small variations in names would not be enough to confidently distinguish a true link from a false link. Another factor in our decision was the large number of names in the sample and complete count data. Although some of the variation is caused by the occasional presence of middle initials, when combined with sex, our given name dictionary contained approximately 1.7 million unique strings.

Given the large number of unique strings, we focused on standardizing strings with a frequency greater or equal to 100 and most of this work dealt with abbreviations and diminutives. Table 2 gives the 30 most frequent male names from our given name dictionary, with the ‘raw’ field containing the original string. The raw string is parsed into three fields (n1, n2, and n3). ‘John W.’ for example, results in n1 = John, n2 = W, and n3 = null. Parsing decisions are based on the presence of a space within the name field and the parsing process also removes non-alpha characters. The table also contains a field for standardized names (n1 standard); ‘Wm’ and ‘Willie’ are standardized as William and ‘Fred’ is standardized as Frederick.

Table 3 lists the most common abbreviations and diminutives found in our data. In addition to the above mentioned examples, here we see Chas standardized as Charles, Joe as Joseph, and so on. The impact of the standardization decisions can also be seen in the table. We use the Jaro-Winkler string similarity algorithm for name comparisons, and the table gives the similarity scores for non-standardized and standardized combinations. For example, combinations like Charlie-Charles, Charley-Charles, Robt-Robert, Thos-Thomas, Saml-Samuel, and Willie-William all receive fairly high similarity scores. The minimum score for these combinations is .910; other given name combinations for verified links with comparable scores would be Ferdinand-Firdnand, Levi-Leevis, Gipson-Gibson, and Shelby-Shelley. Although most of the unstandardized-standardized pairs in the table would emerge as potential links if surname and age were exact matches, they would not ultimately be classified as true links if last name or age were not exact matches. The combinations with the lowest similarity—Jim-James (.720) and Wm-William (.593)—would rarely be classified as true links regardless of similarity for surname and age.

We used Freely Extensible Biomedical Record Linkage (FEBRL) software to construct name and age similarity scores.[[10]](#endnote-10) Records were extracted from our databases based on race and birthplace, with separate files for males, females, and married couples. For our 1870 – 1880 male linked sample, we compare two files, the first consisting of white males, born in Michigan in the 1870 data, and the second file consisting of white males, born in Michigan in the 1880 data. We also use a +/- seven-year age window for comparing records. If a specific record comparison generated scores exceeding preset thresholds, the record pair was written to a results file.

After all files from a given pair of census years have been through similarity score construction, we classified the potential links.[[11]](#endnote-11) A large number of classification techniques exist and their performances vary from domain to domain. In recent years, the use of Support Vector Machines (SVMs) have become an increasingly popular classification choice.[[12]](#endnote-12) The basic concept is that SVMs attempt to maximize separation between the classes, which in this case would be true and false links. SVM construction depends on the existence of training data, which typically consists of a verified set of true and false links.[[13]](#endnote-13) The 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). After SVM construction (which is based on the training data), unclassified records are plotted on this multidimensional space and the end result is a file consisting of potential links and the classifier-produced confidence score. Confidence scores are interpreted dichotomously; a positive score = “true” link and a negative score = “false” link.

A significant feature of SVMs is the absence of diagnostic statistics assessing classifier performance. Classifier evaluation typically depends on the existence of a set of verified links, with analysis focusing on misclassified records (i.e., false positives and false negatives). Unacceptable levels of misclassified records can be dealt with by modifying training data or the set of linkage variables, which in effect redefines the definition of a true link. Another way to interpret the classifier process is to think of an exact match as a single point in a multidimensional space. Given our limited set of linkage variables, the coordinates in this space consist of deviations from the exact match similarity scores for given name, surname, and age. The classifier uses the training data to define the space—in terms of combinations of deviations from exact match scores—that will be interpreted as true links.

At the classifier stage each potential link is evaluated independently, which often results in numerous potential links (from 1880) to a given sample record. We consider these links to be ambiguous, and they are not included in our linked data. Table 4 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, we would say the exact link is probably the correct link. However, we also feel that the probability that it is the correct link is significantly under 95 percent, and using these types of links would introduce an unacceptable error rate.

***The Preliminary Linked Files***

We released preliminary versions of the linked samples based on the procedures outlined above in fall 2008. At that point we were relatively content with the results. Although we could examine the linked samples and identify what appeared to be incorrect links, we also realized that errors were inevitable. A theoretical example would be that there were two individuals alive in 1870 with similar linking characteristics: identical race and birthplace information, and similar if not identical name and age information. The example also assumes that one of these individuals was present in our one-percent sample for 1870 and was thus eligible to be linked to the 1880 census. If the sampled individual died in the ensuing 10 years and the non-sampled record was still alive and enumerated in 1880 with information consistent with the 1870 record, then the non-sampled record would be our link, which would be incorrect. We felt that this scenario was most likely in older age groups (which would be most susceptible to mortality effects), but it could also occur for a variety of reasons: underenumeration; misreporting of name, age, birthplace, or race information; or data entry error.

We also assessed the accuracy of our preliminary samples by comparing them to a set of links for 1870 and 1900 produced by Pleiades Software Development.[[14]](#endnote-14) 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 fairly complex and is based on an additive point system that assesses similarity for individual records. In contrast to our approach, however, they also utilize household and residential information. Given their expanded set of linkage variables, we anticipated the Pleiades linked samples would have a high degree of accuracy and a significantly higher linkage rate. Analysis of the 1870 data showed that Pleiades linked 28.3 percent of the native-born white males compared to a linkage rate of 10.1 percent in our preliminary sample. Again, we expected a significant linkage rate differential and were more interested in the set of 1870 records that were linked in both the Pleiades and MPC linked samples, with the important question relating to whether we consistently linked to the same 1880 record.

Consistency is a relative term, but of the native-born white males present in both linked samples, we disagreed on the 1880 record 1.2 percent of the time. Although we felt that we had the correct link in some of the conflicting cases, we also would have been quite content to interpret all of conflicts as errors in the MPC data if this would imply that we had an error rate of 1.2 percent. However, we could not do this because only 44 percent of our links were present in the Pleiades linked set. Visual examination of the household data for the records that were only present in our linked data disclosed a fair number that we would characterize as ambiguous or likely errors. But the overwhelming majority appeared to be accurate links. The MPC-only links were also likely to be younger individuals, often enumerated as a child in both 1870 and 1880. A review of these links also disclosed that the 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 imprecise parental name, age, or birthplace information.

Table 5 has a few examples from the preliminary release. All of the linked samples list the entire household for specific linked individuals in both the sample year and 1880, and also contain all variables from both the sample year and 1880. For example, the households in the tables show the given name, surname, age, and relationship to head information for both census years. Linked individuals are shown on the same line and ‘Linktype‘ indicates whether a record is a primary link or a household link.[[15]](#endnote-15) The first household listed is the Weeks family. Our preliminary release had Edwin, one year old in 1870 as the primary link. It appears that all other members of the 1870 household were also present in 1880, along with the three youngest children, who were not born as of 1870. The two enumerations of the family are very consistent with a couple of exceptions: given their ages in 1870, both the head and wife’s ages are four years different from their expected age in 1880. The second household in the table is the Underwood family. Our preliminary data had Vander, age 1 in 1870, as the primary link. Irving/Ervin along with their mother Mary is also present in both enumerations. However, Mary’s age is six years greater than expected age in 1880. Another significant difference is the absence of her husband Norman; Mary is listed as a widow in 1880.

The comparison of our links to the Pleiades set indicates that the use of household information would increase linkage rates, but only in cases where household information is consistent. However, calibrating the relationship between agreement or disagreement levels and the classification of potential links is difficult. A linking procedure that expects to see parents with consistent information in two different censuses in order to link children will encounter problems.[[16]](#endnote-16) The process of comparing the MPC and Pleiades linked sets also highlighted the high degree of imprecision in the 19th century census data. For example, our data consists of primary links, and then we proceed to link other household members. By definition, our primary links generally had to be within two years of expected age. But a comparison of the household 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.[[17]](#endnote-17)

Overall, we were fairly confident about the general quality of our linked data after the comparison to the Pleiades links. Again, we had relatively few disagreements with Pleiades in the group of records that we both linked. For the other 56 percent of our links (i.e., the group that Pleiades did not link) we used indirect measures to estimate the error rate. For example, we believe, all things being equal, that differentials in migration rates for groups of linked records is the result of differentials in linkage error rates. In this case, we had two relevant sets of linked records, the set where both Pleiades and MPC linked and the set where only MPC made the link. The migration rate for the former group was 25.4 percent, while the migration rate for the MPC-only links was about 10 percentage points higher. Thus we assume that the error rate for the MPC-only links was approximately 10 percent. Combined with a conservative error rate estimate of two percent for the male links made by both Pleiades and MPC, we arrived at a total error rate of approximately 6 percent.[[18]](#endnote-18)

Although our error rate estimates are relatively imprecise, we felt that the implied error rate was acceptable. However, we also began to think about ways to improve the linkage process for the final data release. A basic issue here was that we had constructed a linkage process that was generally accurate, but was also highly dependent on the precision on the data. More specifically, if the individuals that 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, underenumeration, or misenumeration of linking variable information—we would make an incorrect link if there was another person with similar characteristics in the 1880 complete-count data.

We dealt with this problem by constructing formal measures for the commonness of names. A fairly standard approach in record linkage projects is to construct frequency tables for names, which more or less assesses 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 as opposed to John Smith. But given our minimalist approach to name cleaning and standardization, we run into a problem with minor typos and misspellings which would show up as low frequency names, although many of these names have high similarity to high frequency names (and would in fact 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. We also constructed a 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 would be able to link relatively common names taken from the smaller states of birth—Delaware for example.

Along with the name commonness and birthplace density measures, we also constructed new training data for the final releases. First we selected Pleiades links from the top half of their score distribution and then filtered out links that did not have a family confirmation score. We then matched these links to tables that gave the number of potential links for a given sample record from the preliminary release. If there was only one potential link, then the Pleiades link became a 'yes' in the training data. If there were two potential links, then the Pleiades link was a 'yes' and the non-Pleiades link became a 'no.' If there were three or more potential links, the non-Pleiades links were classified as 'no.' However, in the last example the Pleiades link did not become a yes; we felt these were the type of links that we should not make because the names were either relatively common or the record came from a more populated place of birth.

The initial results were mixed. Given the rules for selecting the training data, records classified as true links all had high levels of similarity. We also saw significant increases in our 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. Again, we felt that a problem with some of our preliminary linked data was the inclusion of these less precise links. But these less precise links were occasionally the mechanism for identifying ambiguous links. We had hoped that the name commonness and birthplace density measures would identify these types of situations, but initially this was not the case.

We considered modifying the training data, but also experimented with dropping linking variables from our models and eventually found an acceptable solution. When we omitted all age information from the classifiers, the resulting linked set had one significant difference from previous results. Indirect measures indicated that this set was generally accurate, but more importantly, the classifier appeared to be reluctant to give positive scores to records with common names, and also appeared to distinguish between similar records depending on whether they came from a high or low density state of birth (although in the absence of age information this model also had the undesirable result of establishing links with low age similarity). Ultimately, we decided to use both models, with releasable links generally defined as records that had one and only one positive link in both models.

We feel that the final released data is more accurate than our preliminary links. In addition, we expanded the number of links for native-born whites. However, we did not add records, and in some cases had a decrease in the number of linked foreign-born whites and African-Americans. The decrease in the number of African American links largely resulted from the decision to use the revised training data, which would rarely result in a link if expected age deviated greatly from actual age. Given the historical background for the enumeration of African Americans in the 19th century, we believe that age precision is generally low for this population.[[19]](#endnote-19)

The immigrant groups also likely suffer from the same problems. But some groups of immigrants also have high levels of name homogeneity, which leads to a high proportion on ambiguous links One way to assess this is to take the product of our name commonness scores for given name and surname. Table 6 gives the distribution of the name score product for males in the 1870 1-percent sample by categories ranging from the least common to the most common. Categories 1 and 2 comprise the bottom two quartiles of the distribution. Categories 3 through 7 represent the top 5 deciles. The rows in the top panel give the distribution for native-born and foreign-born whites and African-Americans by name score categories. The main deviation from average is that both foreign-born whites and African Americans are overrepresented in the top decile of name commonness. The second panel gives the distributions for specific foreign birthplaces. Here we can see that a number of foreign groups—the Irish, Welch, Norwegian, Swede, Scotch and English—are overrepresented in the top decile.

The impact of name commonness on the linkage rates can be seen in Table 7. As expected, all three groups have higher linkage rates for records with less common names compared to those with the most common names. The linkage rate differentials within the ethnic/race groups are also extreme. Taking nothing else into account, native-born whites in the least common name category are approximately 30 times more likely to be linked compared to native-born whites in the most common name category.

However, there is one other basic factor to take into account when evaluating the name commonness scores. Table 8 presents linkage rates by name commonness categories and by birthplace categories, with birthplace rank determined by the proportion of all 1870 males by place of birth. The basic pattern of a higher linkage rate for less common names is evident regardless of birthplace rank. For records in the least common name category, the birthplace categories do not show much difference in linkage rates. But as we move to the more common name categories differentials between less and more populated states become more extreme. For example, only the most common name categories show a linkage rate significantly lower than the average for all records from the small states. The noticeable drop in the linkage rate typically shifts to less common names as we move to the larger states. In the largest states the second least common name category has a linkage rate approximately half that for records in the least common name category. By the time we get to the second most common name category, we have a linkage rate of 0.1%, and we did not link any of the 1984 records from the most common name category for the largest states.

We expected to see this pattern in the linked data. The pattern largely results from our decision not to link in ambiguous cases (i.e., where we had more than one potential link). But by generating name commonness scores for all records, we have also forced the classifiers not to make links in cases where a name is relatively common, even in the absence of competing potential links. And the tables make clear that we do have a biased sample of linked data; in general, we are more likely to link records from the smaller states, and we are more likely to link records with less common names.

We always assumed that we would deal with linkage differentials by constructing weights for the linked records; that all things being equal, linked individuals born in Delaware would have a lower weight than those born in New York. But we also assumed that there would not be an inherent bias between more versus less common names. The construction of the name scores allows us to examine this a fairly straightforward way. Table 9 gives mean occupational score for 1870 males by race/nativity and name category.[[20]](#endnote-20) The first column excludes non-occupational responses and the distribution is relatively flat. We also present the mean occupational score excluding non-occupational responses and farmers, and we also see relatively little variation among name categories.

**Final Release**

Table 10 contains three households from our 1870-1880 male linked sample. The format here is the same as in Table 5; for example, the households in the tables show the given name, surname, age, and relationship to head information for both census years. Linked individuals are shown on the same line and ‘Linktype‘ indicates whether a record is a primary link or a household link. In the first household, the primary link is the third individual (Alva). In this case, there is a high degree of name and age similarity and we link the household members on this basis.

In the second household the primary link is ‘Eddie Cimmerman’ in 1870 and ‘Edward Zimmerman’ in 1880. Although three household members from 1870 are not present in 1880, we also see that there is a high amount of similarity between the other household members despite the different surname spelling. The third household shows an example of a primary link with a relatively rare given name. This contrasts with the household head’s given name information, where we would have difficulty linking two records enumerated as ‘L’ and ‘Lathrop’ in different census years (although once we have established the primary link we will link this individual in the household linking process).

All of the primary links in table 10 appear to be accurate despite some imprecision in the household links. In addition, all of these households were residents of the same state and county in both census years, which makes us very confident that they were linked correctly. But we can also identify links in all of the samples that appear to be incorrect. Although we can never say with absolute certainty that a link is incorrect, the least plausible links typically appear to have a larger number of records in one year that do appear to be the same as those in 1880.

While it is difficult to accurately estimate the number of incorrect links, there are a number of indirect measures that we can use to assess the general accuracy or consistency. For example, we have a linked couples sample for 1870, which consists of couples who were married in 1870 and 1880. In addition, we have male and female only samples where we can identify the linked records that were married and living with their spouse in both 1870 and 1880. We would expect these groups (i.e., the married couple links, and married couples taken from the male linked file and the female linked file) to have comparable characteristics. To the extent that they do not, it should be a reflection of fact that the couples links should be more accurate (all thing being equal) because in addition to name and age information for an individual male or female, we also have the spouse’s given name and age. For 1870 couples where the male is a native-born white, we found 22.3 percent in a different state or different county in the same state 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 respectively.

Finding generally comparable migration rates in the above example speaks more to consistency rather than accuracy. A diagnostic that focuses more on accuracy would be to identify the married males in the linked male file who were married and living with a spouse in both 1870 and 1880. If the male record was also linked in the couples linked file, we can examine whether we linked to the same record from the 1880 file. In other words, we identify male records that are linked in both the male and couples file, and we determine if they are linked to the same household in the 1880 data. And it appears that we rarely have this inconsistency. Of the 3609 males that are in both the male and couples linked files that also have spouses in both 1870 and 1880, we link to different households 8 times.

We can also identify sets of brothers in the 1870 male sample that were enumerated as sons in both years and were living with both parents in both years. Altogether we have 1723 native-born white male links in our 1870 male sample that satisfy this requirement. And we would expect the specific sets of brothers to end up in the same household in 1880. This does not happen for 2.0 percent of the sets. And this works as an indirect error estimate for this group of links. Although some of the consistently linked sets of brothers could be errors, we feel that it would be rare to find inaccurate links among the consistently linked records. Even among the two percent that ended up in one or more different households, we feel that some were accurately linked.

We believe that the estimate for the linked sets of native-born white brothers, along with high levels of consistency for males linked in both the male and couples file is important. Altogether, the sets of brothers and the married males that are also linked in the couples file make up over 25 percent of our 1870-1880 male links. Although we feel that we have higher error rates for both the foreign-born whites and the African American linked populations, we believe that the error rates for the native born linked populations is significantly below 5 percent.

Table 11 gives the number of linked records for the various sample years by nativity/race, and for females by nativity/race and marital status categories. In addition to showing the limitations of some of our samples due to small Ns, we also weight within each of the subgroups. The purpose of the weights is to compensate for under and over representation. As disclosed in table 7 above, in addition to variation in linkage rates among segments of the native-born white population, both the foreign-born and African-Americans have lower linkage rates than the native born whites. We believe this is caused largely by general name and age imprecision, along with greater name homogeneity among the foreign born. But underrepresentation can also be caused by respondent bias. Enumerators went household to household, and we assume the typical procedure was to talk to household heads or the spouses of heads. Less likely was direct communication with unrelated individuals (e.g., boarders, lodgers, and employees). We believe that this group—which was approximately 10 percent of the adult population in the 19th century—is underrepresented in our linked samples because of imprecise name and 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).

There are different strategies for weighting, but the basic approach is to weight by population characteristics in the terminal year. Thus for linked data from the 1850 to 1870 censuses we weight by 1880 characteristics. For the post-1880 sample years we weight by the sample year characteristics. Weights are 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 that was 10 years or older in the 1880 census. However. since we do not have year of immigration information before 1900, for the foreign-born we have to look at the foreign-born in the 1870 census, and apply life tables to estimate how many of these individuals would have still 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 construct an initial weight—which is the inverse of the groups linkage rate—that is used to inflate the linked sample to the actual (or estimated) population totals for all subgroups in the terminal year’s census. We then calculate the specific weights for a number of weighting variables. Again using 1870-1880 males as a example, we first estimate the relationship to head weight, which is calculated as the proportion of the linkable population by relationship-to-head categories divided by the proportions for the linked sample. After this we use the first weight to weight the linked records, and then calculate the proportions for the next weighting variable (in this case individual birthplaces). We repeat this process for 5-year age groups, size of place and occupational categories, with a specific record’s weight getting modified with each iteration.

This process works fairly well if we have enough records. From table 11, however, we can see that some of the sub-groups have small numbers. As a result some subgroups have a large range of weight values, with some records representing a relatively large number of records. We deal with this by imposing a minimum and maximum on the weight for all sub-groups. The minimum is 1/5 of the average weight for the subgroup, with the maximum capped at 4 times the average weight for the subgroup. This has little effect on some of the larger groups; for example, under 1 % of the 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.

Ultimately individual researchers have to decide whether the constructed weights are appropriate for their specific study. Researchers should also be aware of the small number of records for some of the sub-groups listed in table 11, and whether to include these sub-groups in their studies. We hope to alleviate this problem in the future by linking 5 percent samples for 1900 and 1930 to the 1880 complete-count database, which should increase all subgroups listed in table 11 for those years by a factor of five. We also hope to expand the linked samples for the 19th century by a factor of 100. We are currently working on a complete-count database for the 1850 U. S. census. Although this project did not have a record-linkage component, we anticipate developing a grant to use the 1850 data along with complete-count data for 1860, 1870, and 1900. In addition to greatly expanding our current linked samples, using complete-count data for 1850 to 1900 would also allow us to create true longitudinal data, consisting of individuals and their households linked in five different censuses.

1. Generally, eligibility to be linked is dependent on being present in a given sample year and the 1880 complete-count database. For the linked couples, the linkable population is also restricted by the requirement to be married and co-resident in both the sample year and 1880. For females, the linkable population consists of women who did not change their 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. [↑](#endnote-ref-1)
2. “Population Database for the United States,” National Institutes of Health, 5R01-HD039327. 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>). [↑](#endnote-ref-2)
3. Stephan A. Thernstrom, Poverty and progress; social mobility in a nineteenth century city (Cambridge, Harvard University Press) 1964; Michael B. Katz, . The people of Hamilton, Canada West: family and class in a mid-nineteenth-century city (Cambridge: Harvard University Press). 1975.; Peter R. Knights, Yankee destinies: The Lives of Ordinary Nineteenth-century Bostonians (Chapel Hill : University of North Carolina Press), 1991. [↑](#endnote-ref-3)
4. See Joseph Ferrie, “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, 29 (xxxx 1996), 141-156; Avery Guest, “Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century,” *Historical Methods* 20: 63-77. 1987; Richard H. Steckel, “Household Migration and Rural Settlement in the United States, 1850-1860,” Explorations in Economic History, 26 (xxxx 1989), 190-218. [↑](#endnote-ref-4)
5. Hastie, Trevor, Robert Tibshirani and Jerome Friedman, 2nd edition, (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer; Nisbet, Robert, John Elder, Gary Miner, 'Handbook of Statistical Analysis & Data Mining Applications, Academic Press/Elsevier. [↑](#endnote-ref-5)
6. After some experimentation, we used an age window of +/- 7 years of expected age. Thus a record with an age of 23 in 1870 would have an expected age of 33 in 1880 and would be compared to same sex/race/birthplace records between the ages of 26 and 40 in the 1880 complete-count data. [↑](#endnote-ref-6)
7. We anticipated using weights to deal with linkage bias. In contrast to variables like birthplace or age, however, there is no reliable way to construct weights based on migration status or proportions of the population that have transitioned from living with kin to living without kin (or vice versa). [↑](#endnote-ref-7)
8. For the married couples linked file we also use the spouse’s given name, age and birthplace. [↑](#endnote-ref-8)
9. Block, William C. and Dianne L. Star. 1995. Data Entry and Verification in the 1850, 1880 and 1920 Public Use Microdata Samples. *Historical Methods* 28: 63-65; Goeken, Ronald, Cuong Nguyen, Steven Ruggles, and Walter Sargent. 2003. The 1880 United States Population Database. *Historical Methods* 36: 27-34; IPUMS logical edit procedures are discussed at <http://usa.ipums.org/usa/doc.shtml>. [↑](#endnote-ref-9)
10. P. Christen and T. Churches. Febrl - Freely extensible biomedical record linkage (Manual, release 0.3), 0.3 edition, April 2005. [↑](#endnote-ref-10)
11. 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. [↑](#endnote-ref-11)
12. Cristianini, Nello and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. London: Cambridge University Press; Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer-Verlag, New York, 2008. [↑](#endnote-ref-12)
13. 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”). [↑](#endnote-ref-13)
14. <http://www.pleiades-software.com/> [↑](#endnote-ref-14)
15. Primary indicates that this was a link in the classifier process. After identifying primary links we have a process that establishes links within the specific household. For example, we would typically not be able to link someone with a common name like John Smith. However, we might be able to link a Zebediah Smith. And if Zebediah had a co-resident son named John in both census years, then the household linking process will link this John Smith. [↑](#endnote-ref-15)
16. And this problem extends beyond overall linkage rates to issues relating to accuracy. For example, a “true” link can be rejected because of household disagreement, with a “false” linking being accepted due to the absence of household disagreement in the latter case. [↑](#endnote-ref-16)
17. It is difficult to say whether this in 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. [↑](#endnote-ref-17)
18. The estimate for native-born white males was approximately 5 percent. Again, we feel that this estimate is relatively conservative; again, all of the differences in the set that both MPC and Pleiades linked were not errors (i.e., the 2% error rate figure). In addition, the 10% error rate estimate for the MPC-only links might be an overestimate; these records were more likely to younger and to have experienced some sort of disruption in their household structure, characteristics that would also be associated with higher migration rates. Pleiades also assigns linkage points for residing in the same state and county in the two census years, which implies that the MPC-only links would be more likely to be migrants. [↑](#endnote-ref-18)
19. Elo. I. T., and S. H. Preston. 1994. Estimating African-American mortality from inaccurate data. *Demogmphy* 31:427-58; Coale, Ansley J. and Norfleet W. Rives, 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, Vol. 39, No. 1 (Jan., 1973), pp. 3-36. [↑](#endnote-ref-19)
20. Occscore is an IPUMS constructed variable that assigns occupational income scores to specific occupations. See http://usa.ipums.org/. [↑](#endnote-ref-20)