Record linkage at the Minnesota Population Center: new estimates of migration rates, 1850-1930

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

Poster Presentation at 2009 Annual Meeting of the Population Association of America

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. 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 use the 1880 complete-count data and a 1-percent sample of the 1870 population. We recently released preliminary versions of our linked files, and plan on releasing final versions in late 2009.[1]

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 basic 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.[2] More recent linkage studies used "soundex" name indexes to facilitate the linking of individuals who had migrated.[3] But the results here were also mixed; soundex indexes exist for specific states, and searching for migrants then required consulting each state index for a potential match. In addition, the absence of machine readable complete-count data meant that researchers had to consult microfilm of the census manuscripts to locate potential links. Although the results were mixed—the resulting samples were relatively expensive and questions concerning representativeness remained—new developments resulted in continued interest in producing linked datasets. The first would be the availability of the 1880 complete-count database. The complete-count data, along with machine-readable samples for all U.S. censuses since 1850, would allow a fully automated record linkage process that would produce nationally representative linked datasets. And we would accomplish this by taking advantage of developments in record linkage and data mining technology.

The 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 2.5 trillion

comparisons. In an attempt to minimize processing time, we decided to limit 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.[4] Given that there were individuals with incorrectly enumerated or transcribed information in the data, we would lose some potential links because of this decision.

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 a typical record linkage project, we were more concerned with the accuracy and representativeness of our links than with 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. 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 decided to restrict the linkage variables to an individual's given name, surname and age.[5]

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—in the 1880 complete-count data. The original grant proposed identifying duplicate records in the complete-count data based on the core linking variables: given name, surname, race, age and birthplace. Duplicates would be excluded 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 our 1880 data. 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 1870 John Smith (expected age of 30) to 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.

[table 1 here]

Because we expected less than ideal age precision in the data, eliminating 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 these links. What this ultimately meant is that the linked samples would largely consist of "unique" individuals. 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

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.[6] Inconsistent 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 prior to linkage. The name fields, in contrast, receive little processing prior to IPUMS release.

IPUMS data contain 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 first initial. In addition, some enumerators 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 sheer 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 given name 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.[7] 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 2 here]

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 that score close to this 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.

[table 3 here]

We used Freely Extensible Biomedical Record Linkage (FEBRL) software to construct name and age similarity scores. Records were extracted from our databases based on same race and birthplace, with separate files for males, females, and married couples. For example, 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 +/- eight-year age window for comparing records. Thus, a 28-year-old in the 1870 data would have an expected age of 38 in 1880, and would be compared to all records with the same race and birthplace between the ages of 30 and 46 in the 1880 data. If a specific record comparison generated scores exceeding preset thresholds, the record pair was written to a results file.[8]

Classifier

After all files from a given pair of census years have been through similarity score construction, we classify the potential links.[9] 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. 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.[10] 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 the classes (i.e., between the true and false links). After SVM construction, 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 negative score = "false" link.

A significant feature of SVMs is the absence of diagnostic statistics assessing classifier performance. Classifier evaluation 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, which in effect redefines the definition of a true link.

At the classifier stage each potential link is evaluated independently, which often results in numerous potential links (from 1880) to a given sample record. Currently 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.

[table 4 here]

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. In the John Bradley example above, we find four potential links in the space that contains true links.

The Preliminary Linked Files

We released preliminary versions of our linked data in October 2008, with the release of the final versions expected in late 2009. Although we continue to improve our classification process, we believe that improvements will be seen in our linkage rates, as opposed to improvement in error rates. Our refusal to include ambiguous links resulted in a conservative linking approach, which we enhanced through filtering procedures on the classified data. For example, although we did not use the middle name field in the linking process, we delete any link with conflicting middle name information. Visual examination of the respective census households discloses that many of these appear to be correct links, but as a group they also appear to have higher than average error rates.

Table 5 contains three households from our 1870-1880 male linked sample. The format shows 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. After identifying the primary link, we attempt to link the remaining household members. In this case, there is a high degree of name and age similarity and we link the household members on this basis.

[table 5 here]

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).

Primary linked records also receive a weight adjustment. Although the linking process was designed to minimize bias, we assume that some records are more likely to be linked than others. Generally, there is an inverse relationship between the number of individuals with a given place of birth and the linkage rate. From the table below, we can see that New York, Pennsylvania and Ohio have lower linkage rates than Delaware (and others). And this is primarily due to the way we deal with the ambiguous links. For the larger places of birth with higher frequencies we are more likely to find ambiguous links. We also have lower linkage rates in the South, which reflects regional variations in enumeration quality.

[table 6 here]

We also feel that some types of records are also less likely to be linked because of enumerator-respondent bias. For example, unrelated individuals (i.e., individuals unrelated to the head of household like boarders and farm laborers) would be less likely to be have accurate age and birthplace information because it is likely that some enumerators never talked to the unrelated individual (relying on the head or head's spouse for the information). And this could also be generally true for migrants, which would lead to a lower linkage rate for places where migrants were more common like larger cities and specific regions.

We base our weight adjustment on the following variables: age (5-year age groups), region, size of place (categorized city population), relationship to head (dichotomous, either related to the head or not related to head), and birthplace. The calculation is based on the linked sample's terminal year. For example, for the 1880-1900 male linked sample, we calculate the proportion of linked records with a specific characteristic, and divide by the proportion found in the general population for the same characteristic. Ultimately each linked record receives a specific weight adjustment for the five weight adjustment variables listed above. The final weight adjustment is the product of the five individual weight adjustments.

MIGRATION

American social historians generally agree that the United States experienced relatively high migration rates during the 19th century. Disagreements persist, however, over the ultimate magnitude and importance of internal migration. Studies focusing on 19th century American communities showed persistence rates—typically measured as the number of residents identified in successive decennial censuses for a specific locale—at 30 to 50 percent. While the non-persistence rate (i.e., the number of residents not located in the terminal census year) is not necessarily equivalent to the migration rate, these studies implied that a minimum of 50 percent of residents outmigrated over a typical 10-year period.[11]

One issue with the non-persistence rates found in the community studies was that they were significantly higher than migration rates indicated in published 19th century census data. The 1850 census, for example, showed that approximately 25 percent of Americans resided in a state other than their state of birth. The state level migration rates establish a minimum migration rate. Reconciling the state rates with those found in the community studies would require that a significant amount—at least half—of outmigrants would have to migrate to a place within the same state in successive censuses. Since the community studies did not locate destinations of migrants—the persistence and non-persistence rates are based on the non-migrants and not on locating the migrants—there was no way to calculate rates of in-state and out-state migration.[12]

The establishment of persistence rates based on the non-migrants but not the migrants also leads to a general critique of the migration rates implied by the community studies; that the inability to locate an individual in successive censuses for a given locale does not necessarily imply the individual migrated. Obvious issues are the effects of mortality and

underenumeration. Another issue is the basic methodology employed in the community studies. Records were hand-linked, the accuracy of which is dependent on the extent that basic individual-level characteristics were consistently enumerated in the successive censuses. Surviving family units would be easier to hand-link, given information on multiple individuals for cross-verification. But those who had left their family of origin would be difficult to locate in the absence of precise name and age information.

Donald Parkerson's "How mobile were nineteenth-century Americans?" reviews the low persistence rates found in the community studies and compares them to migration rates derived from a 19th century census source. Although the U.S. census did not contain an explicit migration question until 1940, the 1855 New York state census asked individuals how long they had resided in their communities. Parkerson used a sample of the 1855 New York census, controlled for characteristics in the New York communities and those found in a sample of the community studies, and found that the persistence rate in the New York communities was approximately 59 percent compared to a mean of 36 percent in the community studies. He also convincingly demonstrates that the different rates are most plausibly due to the community study methodology; i.e., low persistence rates found in the community studies were mainly caused by the inability to locate individuals who had not actually moved.[13]

Although Parkerson's conclusion's are convincing, his findings did little to end the desire of researchers to construct linked data sets to explore issues related to migration. One reason is that in addition to specific rates, researchers were also interested in identifying specific characteristics of migration. And the tone was set for this interest by Frederick Jackson Turner's 1893 essay concerning the closing of the American frontier. Turner argued that the nineteenth century was a period of high migration, much of it motivated by the availability of land on the frontier. Turner also characterized westward migration as an enticing option for laborers in the more urbanized east, but that the closing of the frontier would result in lower migration in the future.[14]

Many of the community studies explicitly examined Turner's hypotheses, but given their methodological deficiencies, researchers continued to pursue better data sources—essentially nationally representative samples that would include both non-migrants and migrants. The ability to construct these data would be enhanced by the existence of indexes listing all heads of household for various U.S. censuses. The availability of computers and the ability to search machine readable versions of the indexes along with the availability of nationally representative samples of the nineteenth-century censuses would lead to further attempts to produce linked datasets.

One significant attempt to construct a representative linked data set was by Richard Steckel, who linked nearly 1600 households in the 1850 and 1860 censuses. The procedure was to identify a random sample of households in 1860 that had at least one child over the age of 9. The birthplaces of children 10 and older provided evidence of the household's state of residence 10 years prior, which was necessary to narrow the search for the household in the 1850 census. Regarding overall migration rates, Steckel found that approximately 30 percent of his households migrated between 1850 and 1860, a

figure generally consistent with Parkerson's conclusions from the 1855 New York state census.[15]

Given that Steckel relied on the presence of children to successfully link households, his household links were undoubtedly accurate. Representativeness was still an issue, however, in that his data underrepresented younger adults and unrelated individuals, who would be expected to be more likely to migrate. Joe Ferrie attempted to correct for this bias by taking a sample of individuals in the 1850 PUMS, and linking them to an index of household heads and unrelated individuals in the 1860 U.S. census. The result, according to Ferrie, produced "longitudinal data more representative of the antebellum U.S. economy than samples linked backwards, and capturing the experiences of younger, more footloose, less established individuals that those samples contained." Ferrie found a migration rate of 47 percent for native-born white males found in both the 1850 and 1860 censuses. His estimate is higher than Steckel's, which Ferrie credits to the characteristics of Steckel's data. While the 47 percent figure is quite a bit lower than the various community studies, it is also somewhat higher than Parkerson's estimates from the New York state census.[16]

Figure 1 gives migration estimates for our linked data, with migration defined as residing in either a different state or a different county within the same state. All data points are in reference to 1880; the figure for 1850-1880 indicates that approximately 58 percent of linked males between those census years were living in a different state or different county within the same state in 1880. Figure 1 also shows that the migration declined to 51 percent for 1860-1870 and 36 percent for 1870-1880.

[Figure 1 here]

None of the rates in Figure 1 are directly comparable to those of other researchers cited above. Ferrie, for example, using a nationally representative sample, estimates the male migration rate for 1850-1860 at 47 percent. The only ten-year period in Figure 1 is 1870-1880, where we find a migration rate significantly lower at 36 percent. However, it is possible that migration was declining in the decades following the Civil War. Although Turner would not declare the frontier closed until 1893, by the 1870s the frontier consisted of the plains and mountain states, which turned out to be less hospitable to densely settled farm communities. We can also compare Ferrie's migration rate for 1850-1860 to our rate for 1850-1880, which is 57 percent. The issue here, all things being equal, is whether we would expect an additional 10 percent of a linked population to migrate over the next twenty years. Although we cannot directly examine this issue, the answer would depend on the rate of return migration; some of the 1850-1860 migrants would not be migrants for the 1850-1880 period if they returned to their 1850 state and county of residence. Among the migrants in our data, approximately 10 percent migrate to their state of birth in the pre-1880 census years (e.g., someone born in Ohio, enumerated in Illinois in 1870, and then enumerated in Ohio in 1880). Although we cannot tell if they returned to their county of origin, we can also assume that return migration would be greater for instate compared to outstate migrants. And, generally, return migration would deflate the expected cumulative effects on migration rates over

increasing time periods.

The view that migration began to decline at some point in the decades following the Civil War is generally consistent with age-standardized lifetime migration rates for the 19th century, which measure migration on the basis of residing in a state other than state of birth. Although the lifetime approach does not capture instate migration, it does capture the high rates of migration in 19th century America. Kelly and Ruggles show that slightly less than half of native-born males age 50-59 were lifetime migrants in the 1850 through 1880 U.S. censuses. The percentage declined after 1880, reaching a low point of approximately 33 percent in 1940. Their figures are for 50-59 year olds, and since younger adults are more likely to migrate, it is probable much of the migration for this group occurs 20-30 years prior to census enumeration. This would indicate that the decline in the lifetime migration rates which they place between 1880 and 1900, was actually occurring between 1860 and 1880. But we can also see in Figure 1 that geographic mobility continued to be a characteristic of the American population. Approximately 43 percent of our linked males for the years 1880 to 1900 migrated. The migration rate for 1880 linked males increases in subsequent census years, with 60 percent of males found in the 1930 census residing in a different state or county.[17]

Although our linked data can only indirectly address the specific timing of a decline in 19th century migration rates, we can examine the characteristics of 19th century migrants to see if they changed over time. One issue is the basic nature of 19th century migration. Although Turner emphasized the closing of the frontier and predicted a decrease in migration in the future, others felt that the rural nature of American migration was somewhat exaggerated Given the growth of American cities during the later part of the 19th century, rural to urban movement must have constituted a fair amount of internal migration.[18]

The migration rates given in Figure 1 are based on defining migration as residence in a different state or different county in the same state. We have also calculated the distance from the center point of county of origin to the center point of county of destination, and Table 7 gives the distribution of migration distance (in miles) for interstate and intrastate migrants for 1870-1880 and 1880-1900. For example, over 40 percent of intrastate migrants travel less than 30 miles compared to less than 7 percent of the interstate migrants in 1870-1880. In contrast, for the same set of census years, over 60 percent of interstate migrants travel more than 250 miles compared to approximately 2 percent of the intrastate migrants. The bottom panel, which gives figures for 1880-1900, show little change in the given categories. The mean and median miles migrated also show little change; for example, interstate migrants moved, on average, 15 more miles in the latter period, but the median actually decreased.

[Table 7 here]

Obviously, there are physical constraints on the maximum distance an intrastate migrant can travel, and this is also ultimately true for interstate migrants as well. For the remainder of this paper, we will define migrants as those that moved at least 30 miles and

will not distinguish between intrastate and interstate migrants. Depending on the specific topic, other researchers may come up with alternate definitions.[19]

Table 8 shows region of origin and destination for male migrants in the 1870-1880 linked data. The row percentages equal 100 percent; e.g., 63.9 percent of male migrants residing in New England in 1870 were also found in New England in 1880, 12.8 percent in the Middle Atlantic region, and so on. The bolded cells represent the percentage of male migrants who did not leave their region of origin. The lowest total, for example, is for the East North Central region, where slightly more than 50 percent of male migrants that resided there in 1870 remained in the region ten years later. Although we can say that most migrants remained in their region of residence over the ten-year period, the table also discloses that male migrants tended to move west. This can be seen by comparing the N for the row, which is the total residing in a given region in 1870, to the N for the column, which gives the total residing in a given region in 1880. Eastern regions all experienced a net decline in population due to internal migration, with the western regions experiencing a net gain.[20]

[Table 8 here]

The bottom panel in table 8 gives a similar origin/destination table by population totals. The top row gives the distribution for residents of rural places in 1870. Over 80 percent of these individuals were also found in rural places in 1880, with 8.9 percent in places with a population between 2500 and 50,000, 3.2 percent in metro areas outside of central cites, and 5.8 percent in central cities in metro areas. Although rural places experienced a net decline of 2.6 percent from 1870 to 1880, we can also say that rural to rural migration was the most common type of move; 62.7 of all 1870-1880 migrants experienced this type of move. In contrast, metropolitan areas did see a net increase due to internal migration between 1870 and 1880, although the size of this type of migration was much smaller than the rural to rural migration.[21]

Table 9 gives similar numbers for 1880-1900. The regional migration flows here generally resemble those seen in the previous table. Western regions continued to experience growth through internal migration, although the West North Central region's growth declined. But the distribution of migrants by type of place does show a significant amount of change. Although rural to rural migration still represents the largest single cell, it declines here to 45.5 percent of all internal migrants. Although only 5.8 percent of rural migrants ended up in metropolitan central cities between 1870 and 1880, the figure for 1880 to 1900 is 17.2 percent. And residents of other types of places were also increasingly likely to settle in big cities. The percent increase/decrease figures also show an interesting contrast with the 1870-1880 figures. Rural places experienced a much larger decline through internal migration compared to the earlier period, and the other three categories experienced net gains, with the biggest increase in the metropolitan central cities.

[Table 9 here]

Although comparisons between the two periods is complicated because of the 10 versus 20 year period, it does appear that the nature of internal migration began to change following 1880. Where Kelly and Ruggles characterized 19th century as predominantly rural in nature, they were also looking at the lifetime migration of men age 50 to 59. Since migration is more likely to occur at younger ages, it is probably that the shift to urban destinations which they identify in the 1930 census, was in fact occurring by the last two decades of the 19th century.

Table 10 gives logistic regression results for the determinants of migration for the 1870-1880 linked males. All independent variables reflect 1870 status. We limit the population to native-born whites in order to assess the impact of lifetime migration prior to 1870. We also restrict the age group to 30 to 44 year olds in order to evaluate the wealth information contained in the 1870 census. Coefficients are given in the first column and the change in the odds ratios is given in the last column (Exp(B)).

[Table 10 here]

One somewhat surprising result is the relative lack of importance for numerous independent variables. Personal property and combinations of marital status and children appear to have little effect on the decision to migrate. Living in western regions and rural areas increased the probability of migrating, although the coefficient is not significant at the 0.05 level for rural places. Farmers were less likely to migrate, in contrast to farm laborers and those in white collar occupations, although none of the results for the non-farmer categories are significant. But there appear to be a couple of characteristics that were very influential in the decision to migrate. Men who did not own any real estate were much more likely to migrate compared to other categories of real estate wealth. This was also true of those who had already migrated prior to 1870 (measured by not residing in their state of birth).

Although the results here are preliminary, it does appear that the absence of real estate wealth was a significant determinant in the decision to migrate. This mechanism could work in a variety of ways. Men who wanted to pursue a livelihood connected to land ownership would be likely to move to areas where land was easier to obtain, and this would be true of frontier areas. It is also possible that land was a relatively non-liquid form of wealth, and that owning land, at least for some men, was an impediment to migration. Men who did not own land were in a generally better position to take advantage of opportunities that migration provided.

Table 11 gives logistic regression results for native-born whites, age 5 to 14 in 1880, who were living with fathers who were also native-born white. Here we are looking at the determinants of migration between 1880 and 1900. Independent variables are similar to those in Table 10, although we do not have wealth variables after 1870. We also use the father's 1880 occupation and whether the father was a lifetime migrant in 1880 (and all variables reflect 1880 characteristics).

One result in table 11 that is consistent with table 10 is the effect of region, where living in the western part of the county in 1880 increased the probability of migration over the 20 year period. The sons of farmers were less likely to migrate, but having a father with a white collar occupation significantly increased the likelihood of migration. Residing in rural areas also had a strong positive effect on migration. Having a father who was a lifetime migrant was also significantly positive, but not as strong as the child being a lifetime migrant (meaning that the household had crossed state lines between the child's birth and the census enumeration).

CONCLUSION (tentative)

The findings above concerning lifetime migration status are interesting. All we can say definitively about lifetime migrants is that they moved and crossed state boundaries at some point in their lives. Although it would be difficult to measure the extent that non-lifetime migrants had more extensive kinship and social networks compared to lifetime migrants, we believe that this is probable, and played a large role in subsequent migration decisions.

We also feel that the basic characteristics of migration was changing in the decades following the Civil War, with the (gradual) closing of the frontier playing a role. The draw of relatively inexpensive undeveloped land in frontier areas decreased over time, and migrants increasingly moved to cities. The results for net increase/decrease by type of place indicates that migrants generally were increasingly likely to move from rural to urban places, and the logistic regression results indicate that this was definitely true for the youngest generation in 1880.

[1] "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).

[2] 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.

.

[3] 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.

[4] After some experimentation, we used an age window of +/- 8 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 41 in the 1880 complete-count data.

[5] For the married couples linked file we also use the spouse's given name, age and birthplace.

[6] 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*. Forthcoming; IPUMS logical edit procedures are discussed at http://usa.ipums.org/usa/doc.shtml.

[7] Only 19,688 of the unique strings have a frequency greater or equal to 100. In contrast, almost 1.4 million of the strings have a frequency of one.

[8] P. Christen and T. Churches. Febrl - Freely extensible biomedical record linkage (Manual, release 0.3), 0.3 edition, April 2005.

[9] 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 (double metaphone).

[10] 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

[11] Thernstrom; Katz; Knights.

[12] Kelly and Ruggles

[13] Donald H. Parkerson, "How Mobile Were Nineteenth-Century Americans?," Historical Methods, 15 (Summer 1982), 99-110.

[14] Frederick Jackson Turner, speech before the American Historical Association, July 12, 1893, in Proceedings of the Forty-first Annual Meeting of the State Historical Society of Wisconsin (Madison, 1894), 79-112.

[15] Richard Steckel.

[16] Ferrie, Joseph. 1996. 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: 141-156.

[17] Kelly and Ruggles.

[18] Kelly and Ruggles.

[19] In addition, some of the short distance movers show up as migrants because of political reorganization or boundary changes at the county level. For example, in 1870 the city of St. Louis was located in St. Louis county Missouri. In 1877 the city of St. Louis became an independent city (or county equivalent). Thus residents of St. Louis in the 1870-1880 linked data will show up as intrastate migrants because they have a different county code in the two years. Because 'false' migrants like St. Louis residents typically have relatively low miles migrated values, we decided to only consider those who migrated at least 30 miles to be 'true' migrants for this paper.

[20] The regions consist of:
New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont.
Middle Atlantic: New Jersey, New York, Pennsylvania.
East North Central: Illinois, Indiana, Michigan, Ohio, Wisconsin.
West North Central: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota.

South Atlantic: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia.
East South Central: Alabama, Kentucky, Mississippi, Tennessee.
West South Central: Arkansas, Louisiana, Oklahoma/Indian Territory, Texas.
Mountain: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming .
Pacific: California, Oregon, Washington

[21] Metropolitan areas are counties or combinations of counties centering on a substantial urban area. Prior to 1950, the Census Bureau did not define metropolitan areas. However, the IPUMS constructed metropolitan areas for 1850-1880 and 1900-1920 using the 1950 census definition of Standard Metropolitan Area. http://usa.ipums.org.

Table 1. Frequency of white males named John Smith, by State of Birth (1880 complete-count data)

AGE	New York	Ohio	Maryland	Mississippi	Iowa	Oregon
20	45	40	7	7	5	1
21	56	46	10	9	11	
22	59	44	11	5	17	1
23	54	39	14	1	16	1
24	59	45	9	6	11	1
25	67	31	12	7	8	
26	42	40	9	5	4	1
27	42	37	5	4	7	1
28	53	45	8	4	3	1
29	34	34	12		5	1
30	51	47	7	6	1	1
31	28	23	1	4	2	1
32	44	32	11	3	8	
33	25	26	6	1	1	
34	25	24	5	6	3	
35	29	33	13	4	2	
36	33	40	5	2	1	
37	29	21	8	4	1	
38	41	25	6	1		
39	24	32	2		4	
40	55	31	6	1		
41	23	11	8			
42	24	15	7		1	
43	30	15	6			
44	11	13	1		1	
45	35	18	5	4		
46	29	22	4			
47	30	13	3			
48	28	14	5	1	1	
49	24	22	2	1		
50	34	24	6	1	1	

Table 2. Name standardization

Raw string	n1	n2	n3	n1 standard	Frequency
JOHN	JOHN				3297148
GEORGE	GEORGE				1311868
WILLIAM	WILLIAM				1161737
HENRY	HENRY				1156464
JAMES	JAMES				792520
CHARLES	CHARLES				591368
THOMAS	THOMAS				474816
JOSEPH	JOSEPH				458782
FRANK	FRANK				382949
PETER	PETER				382598
EDWARD	EDWARD				298570
ROBERT	ROBERT				241233
SAMUEL	SAMUEL				217773
DAVID	DAVID				189674
JACOB	JACOB				186299
ALBERT	ALBERT				172205
DANIEL	DANIEL				157188
WM.	WM			WILLIAM	152013
MICHAEL	MICHAEL				147171
ANDREW	ANDREW				137175
PATRICK	PATRICK				136865
WILLIE	WILLIE			WILLIAM	128333
RICHARD	RICHARD				123986
LOUIS	LOUIS				115083
JOHN W.	JOHN	W			113715
HARRY	HARRY				111693
GEORGE W.	GEORGE	W			105594
FRED	FRED			FREDERICK	102832
WILLIAM H.	WILLIAM	H			102777
WALTER	WALTER				99388

Table 3. Common male abbreviations and diminutives.

Raw string	n1	n1 standard	Frequency	Jaro-Winkler
WM.	WM	WILLIAM	152013	0.593
WILLIE	WILLIE	WILLIAM	128333	0.910
FRED	FRED	FREDERICK	102832	0.870
CHAS.	CHAS	CHARLES	65007	0.900
JOE	JOE	JOSEPH	63356	0.867
GEO.	GEO	GEORGE	47135	0.867
CHARLEY	CHARLEY	CHARLES	46206	0.943
THOS.	THOS	THOMAS	42224	0.922
SAM	SAM	SAMUEL	39799	0.867
CHARLIE	CHARLIE	CHARLES	38491	0.943
EDDIE	EDDIE	EDWARD	36106	0.760
ROBT.	ROBT	ROBERT	28990	0.922
ALEX	ALEX	ALEXANDER	26199	0.870
JAS.	JAS	JAMES	25740	0.893
BEN	BEN	BENJAMIN	24261	0.833
JNO.	JNO	JOHN	23679	0.825
TOM	TOM	THOMAS	21772	0.850
JIM	JIM	JAMES	20176	0.720
SAML.	SAML	SAMUEL	14691	0.922
BENJ.	BENJ	BENJAMIN	14298	0.883
JOS.	JOS	JOSEPH	12072	0.867

Table 4. Potential links and confidence scores

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
john	bradley	john	clardy	25	34	-1.047429204

Table 5. Linked record example

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 6. Linkage Rates

State or Country of Birth	Linkage Rate
Alabama	6.4
California	12.5
Connecticut	16.9
Delaware	18.6
Georgia	6.9
Indiana	10.3
Louisiana	7.8
Michigan	15.6
New Hampshire	17.3
New York	6.3
Ohio	8.7
Pennsylvania	7.4
Utah	18.0
Virginia	7.7
Canada	10.1
Norway	3.6
Sweden	3.6
England	9.0
Scotland	8.8
Ireland	2.1
Czechoslovakia	10.0
Germany	5.2
Total	8.6
Relationship to Head	
Related to Head	9.0
Not Related to Head	5.1
Total	8.6
Region of Residence	
New England	13.7
Middle Atlantic	7.4
East North Central	9.1
West North Central	8.7
South Atlantic	9.3
East South Central	6.8
West South Central	6.7
Mountain	8.3
Pacific	7.8
Total	8.6

Population of place of residence	Linkage Rate
Under 1,000 or unincorporated	9.1
1,000-2,499	8.4
2,500-3,999	8.2
4,000-4,999	7.8
5,000-9,999	8.2
10,000-24,999	8.1
25,000-49,999	6.9
50,000-74,999	7.2
75,000-99,999	5.6
100,000-199,999	8.0
200,000-299,999	6.8
300,000-399,999	4.1
600,000-749,999	6.3
750,000-999,999	4.3
Total	8.6
Age	
0-4	10.1
5-9	9.1
10-14	8.4
15-19	7.5
20-24	7.0
25-29	7.7
30-34	7.8
35-39	7.7
40-44	8.6
45-49	8.9
50-54	9.2
55-59	11.1
60-64	11.1
65-69	11.9
70-74	10.4
75-79	7.3
80-84	3.5
85-89	2.2
Total	8.6

Figure 1. Male migration rates, 1850 – 1930.

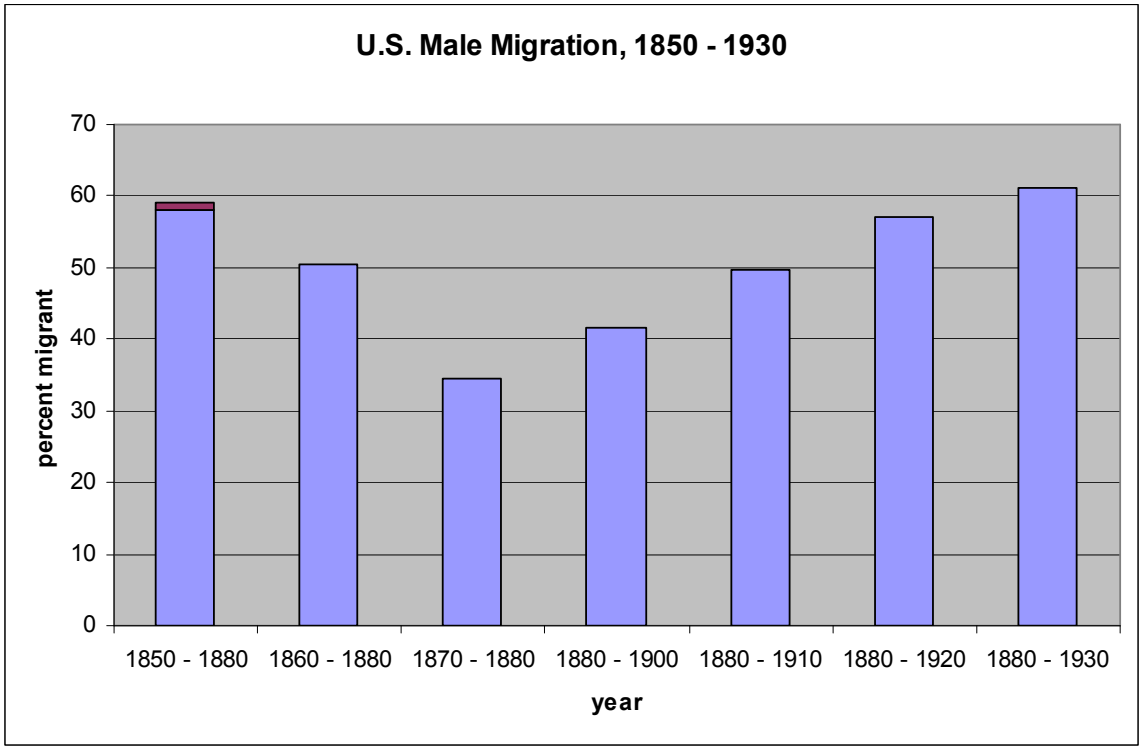
U.S. Male Migration, 1850 - 1930
70
60
50
40
30
20
10
0
percent migrant
1850 - 1880
1860 - 1880
1870 - 1880
1880 - 1900
1880 - 1910
1880 - 1920
1880 - 1930
year

Table 7. Distance migrated for linked 1870-1880 and 1880-1900 linked males

1870-1880 distance	instate migrants	outstate migrants
1 to 29 miles	43.5	6.4
30 to 99 miles	36.2	12.8
100 to 249 miles	17.8	19.7
ge 250 miles	2.4	61.0
mean	63	495
median	37	345

1880-1900 distance	instate migrants	outstate migrants
1 to 29 miles	39.6	6.3
30 to 99 miles	40.6	11.5
100 to 249 miles	18.0	20.0
ge 250 miles	1.8	62.2
mean	63	510
median	36	341

Table 8

Table 8a. Region of origin and destination, male migrants 1870-1880.

	New England	Middle Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Mountain	Pacific	N
New England	**63.9**	12.8	9.5	5.3	1.8	0.8	0.3	1.3	4.5	399
Middle Atlantic	6.3	**54.6**	19.9	10.0	3.1	0.8	1.0	1.5	2.8	617
East North Central	2.3	6.0	**49.8**	32.0	2.6	2.5	1.4	1.2	2.2	1038
West North Central	1.8	4.1	16.4	**59.7**	2.5	2.3	4.9	4.1	4.3	610
South Atlantic	1.2	7.8	4.6	4.4	**67.7**	6.9	6.5	0.4	0.7	758
East South Central		0.7	7.7	7.2	8.2	**56.7**	18.8	0.2	0.5	559
West South Central	0.7	1.3	1.7	2.7	10.0	10.6	**72.8**	0.3		301
Mountain	1.7	3.4	12.1	6.9			3.4	**65.5**	6.9	58
Pacific	8.5	2.3	9.3	3.9		2.3	1.6	4.7	**67.4**	129
N	352	547	880	869	657	452	429	100	183	4469

	1870 N	1880 N	% increase/decrease
New England	399	352	-11.8
Middle Atlantic	617	547	-11.4
East North Central	1038	880	-15.2
West North Central	610	869	+42.5
South Atlantic	758	657	-13.3
East South Central	559	452	-19.1
West South Central	301	429	+42.5
Mountain	58	100	+72.4
Pacific	129	183	+41.2

Table 8b. 1870 - 1880

	rural (city population lt 2500)	2500 to 49999	metro, fringe	metro, central city	N
rural (city population lt 2500)	82.0	8.9	3.2	5.8	3419
2500 to 49999	50.1	20.7	8.8	20.5	547
metro, fringe	59.7	13.6	6.5	20.1	154
metro, central city	46.1	17.5	5.2	31.2	349
	3330	500	187	452	4469

	1870 N	1880 N	% increase/decrease
rural (city population lt 2500)	3419	3330	-2.6
2500 to 49999	547	500	-8.6
metro, fringe	154	187	+21.4
metro, central city	349	452	+29.5

Table 9a. Region of origin and destination, male migrants 1880-1900.

	New England	Middle Atlantic	East North Central	West North Central	South Atlantic	East South Central	West South Central	Mountain	Pacific	N
New England	**54.6**	23.0	8.1	4.5	0.6		0.3	2.7	6.3	335
Middle Atlantic	6.1	**64.7**	12.8	6.2	1.8	0.6	1.0	3.3	3.4	1080
East North Central	1.3	7.1	**51.3**	23.4	1.3	2.3	3.2	4.5	5.6	1413
West North Central	0.7	4.5	12.4	**51.8**	0.8	1.1	10.0	8.8	10.0	1040
South Atlantic	0.7	10.6	4.7	4.2	**62.1**	7.7	6.0	2.0	1.9	697
East South Central	0.9	1.5	8.1	8.2	4.1	**50.5**	24.5	0.9	1.3	681
West South Central	0.2		1.5	2.9	2.6	7.3	**83.0**	2.0	0.4	454
Mountain	0.9	5.3	7.0	7.9	3.5	1.8	1.8	**59.6**	12.3	114
Pacific	2.2	1.4	7.2	4.3			0.7	7.2	**77.0**	139
N	290	1016	1132	1065	524	483	750	307	386	5953

	1880 N	1900 N	% increase/decrease
New England	335	290	-13.4
Middle Atlantic	1080	1016	-5.9
East North Central	1413	1132	-19.9
West North Central	1040	1065	+2.4
South Atlantic	697	524	-24.8
East South Central	681	483	-29.1
West South Central	454	750	+65.2
Mountain	114	307	+69.3
Pacific	139	386	+77.7

Table 9b. 1880 – 1900

1880-1900 city population

	rural (city population lt 2500)	2500 to 49999	metro, fringe	metro, central city	N
rural (city population lt 2500)	60.0	16.9	5.9	17.2	4511
2500 to 49999	26.2	23.6	7.8	42.4	768
metro, fringe	36.4	18.5	10.8	34.4	195
metro, central city	22.2	20.6	9.8	47.4	481
	3087	1077	396	1395	5955

	1880 N	1900 N	% increase/decrease
rural (city population lt 2500)	4511	3087	-31.6
2500 to 49999	768	1077	+40.2
metro, fringe	195	396	+103.1
metro, central city	481	1395	+190.0

Table 9. Logistic regression results

Logistic Regression Results; Native Born, White Males, age 30 to 44 in 1870
(Dependent variable = migrated between 1870 and 1880)

	B	S.E.	Wald	df	Sig.	Exp(B)
Real Estate			31.502	3	.000	
$0	.584	.108	29.044	1	.000	1.793
$1 to $999	.003	.133	.000	1	.983	1.003
$1000 to $2999	-.116	.121	.918	1	.338	.890
($3000+)						(reference)
Personal Property			.212	3	.976	
$0 to $99	.001	.119	.000	1	.993	1.001
$100 to $299	-.053	.124	.181	1	.671	.948
$300 to $999	.029	.102	.084	1	.772	1.030
($1000+)						(reference)
West	.207	.081	6.621	1	.010	1.230
(not West)						(reference)
City Population			4.480	3	.214	
0 to 2499	.221	.129	2.924	1	.087	1.248
2500 to 49999	-.171	.157	1.179	1	.278	.843
Metro, Fringe	.056	.221	.065	1	.799	1.058
(Metro, Central City)						(reference)
Age			4.725	2	.094	
30 to 34	.133	.082	2.590	1	.108	1.142
35 to 39	.058	.083	.490	1	.484	1.060
(40 to 44)						(reference)
Occupation			12.863	4	.012	
Farmer	-.376	.114	10.942	1	.001	.687
Farm Laborer	.129	.174	.553	1	.457	1.138
White Collar	.220	.152	2.088	1	.148	1.246
Sales/Craft/Operatives	-.077	.120	.418	1	.518	.926
(Laborers)						(reference)
Lifetime Migrant	.476	.064	55.139	1	.000	1.610
(Not Lifetime Migrant)						(reference)
Marital Status/Children			.184	2	.912	
Single	-.019	.128	.022	1	.883	.981
Married, w/o Children	-.023	.145	.026	1	.872	.977
(Married w/ Children)						(reference)
Constant	-1.251	.153	66.463	1	.000	.286

Table 11. Logistic regression results: native born, white males, age 5 to 14 in 1880.
(Dependent variable = migrated between 1880 and 1900)

Table 11. Logistic regression results, Native born, white males, age 5 to 14 in 1880. (Dependent variable = migrated between 1880 and 1900)

	B	S.E.	Wald	df	Sig.	Exp(B)
West	.253	.051	24.318	1	.000	1.288
(East)						#####
Age 5 - 9	.069	.041	2.878	1	.090	1.071
(Age 10 -14)						#####
Occupation			45.509	4	.000	
Farmer	-.338	.086	15.612	1	.000	.713
Farm Laborer	-.462	.230	4.029	1	.045	.630
White Collar	.457	.117	15.396	1	.000	1.580
Sales/Craft/Operative	.154	.105	2.156	1	.142	1.167
(Laborers						#####
Father = lifetime migrant	.147	.046	10.184	1	.001	1.158
(Father not lifetime mig.)						#####
Lifetime Migrant	.445	.069	41.864	1	.000	1.560
(Not lifetime mig)						#####
City Population			39.573	3	.000	
0 to 2499	.606	.101	35.972	1	.000	1.832
2500 to 49999	.317	.124	6.552	1	.010	1.373
Metro, Fringe	-.219	.170	1.650	1	.199	.804
Metro, Central City						#####
Constant	-.286	.115	6.132	1	.013	.751