CARRA Working Paper Series

Working Paper #2016-05

# Playing with Matches: An Assessment of Accuracy in Linked Historical Data

Catherine G. Massey
U.S. Census Bureau

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: June 2016

# Playing with Matches: An Assessment of Accuracy in Linked Historical Data

Catherine G. Massey, U.S. Census Bureau

**Abstract**

This paper evaluates linkage quality achieved by various record linkage techniques used in historical demography. I create benchmark, or truth, data by linking the 2005 Current Population Survey Annual Social and Economic Supplement to the Social Security Administration's Numeric Identification System by Social Security Number. By comparing simulated linkages to the benchmark data, I examine the value added (in terms of number and quality of links) from incorporating text-string comparators, adjusting age, and using a probabilistic matching algorithm. I find that text-string comparators and probabilistic approaches are useful for increasing the linkage rate, but use of text-string comparators may decrease accuracy in some cases. Overall, probabilistic matching offers the best balance between linkage rates and accuracy.

**Key words:** Record linkage, historical demography, microdata, censuses

## 1. Introduction

Longitudinal data facilitate research of life course phenomena, social change, and mobility; however, historical person-level longitudinal survey data are virtually nonexistent in the U.S. To create longitudinal data, historians use personally identifiable information (PII) to link person-records across existing data sources.[1] Ferrie (1996) created one of the first national, linked samples of the 1850 and 1860 censuses, relying on phonetic codes of first and last names, age, and place of birth to establish links.

Since Ferrie (1996), there have been several improvements in historical linking methods. These include use of text-string comparators as well as sophisticated record linkage algorithms that employ probabilistic matching techniques and machine learning. The Minnesota Population Center (MPC), for instance, uses the Freely Extensible Biomedical Record Linkage (FEBRL) software and Support Vector Machine (SVM) software, which employs comparison routines to score the similarity of two records, to link individuals across the 1850-1930 censuses (Ruggles 2006, 2011). They also account for name and birthplace commonness, essentially incorporating the probability of a successful match (Goeken et al. 2011). Similarly, Antonie et al. (2014) and Mill and Stein (2013) score matches and incorporate the probability of a true match into their matching algorithm. Despite recent advancements, little work evaluates whether these techniques improve the quality of linked data in historical contexts.

Evidence suggests the links established across historical datasets are accurate. For instance, Wisselgren et al. (2014) link 1890 and 1900 Swedish census data and compare these links to digitized parish registries to evaluate accuracy. They confirm 97.5 percent of linked men

[1] PII refers to any information that can identify an individual such as name, date of birth, and birthplace.

from the censuses with parish records (Wisselgren et al. 2014, 148).[2] To assess accuracy of the publicly available linked U.S. decennial census data produced by the MPC, Goeken et al. (2011) examined the 1870 and 1880 U.S. linked samples. Using household-composition information of white, native-born married couples and young brothers, they determine the percentage of erroneous links is small, at two percent or less.[3] These conclusions, however, are specific to the data sources and linkage techniques used in each analysis.

In this paper, I assess the accuracy of data produced from historical linking techniques by starting with Ferrie's (1996) standard record linkage technique and systematically incorporating recent innovations. First, I establish a benchmark set of "correctly" linked records using the 2005 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) linked to the Social Security Administration's (SSA) Numeric Identification System, or Numident, by Social Security Number (SSN). Next, I link the CPS ASEC to the Numident using Ferrie's (1996) standard record linkage technique (exact matching on phonetically coded first and last name, age, and place of birth) and compare these links to the benchmark links. After establishing baseline linkage and accuracy results, I then propose and test methods that may improve linkage rates and accuracy beyond the baseline. In particular, I assess the value added from using text-string comparators to compare first and last names, editing date of birth, and using probabilistic matching techniques. Since the Numident data contain date of death, I also explore the effects of death between survey years on linkage rates and accuracy.

[2] Wisselgren *et al.* employ name standardization techniques similar to those used by the MPC (see Vick and Huynh 2011). They link records using standardized names, parish of birth, year of birth and residence to link records across censuses. After editing names and using household information in the match, the percent of confirmed records increases to 98.3 percent.

[3] Goeken et al. (2011) compare households of married males in the 1870 census to their household in 1880 and determine that only 8 out of 3,609 males were linked to different households in 1880 (Goeken et al. 2011, 12). They also look at brothers in the 1870 census who were young enough to have been enumerated with their parents in 1880. They find only 2.0 percent of brothers were linked to the wrong household (Goeken et al. 2011, 12).

This paper has four major findings that contribute to the literature on historical record linkage. First, depending on accuracy of the linkage keys, accuracy of the linkage can be low. Inconsistencies in the measurement of age across data sources, in particular, greatly affects accuracy. Second, attrition resulting from deceased individuals is often cited as an explanation for low match rates (Guest 1987); however, I find that mortality may actually inflate linkage rates by shrinking the pool of potential matches. Third, text-string comparators increase the linkage rate, but may do so at the cost of accuracy. Last, I determine probabilistic matching techniques combined with string comparators offer the best balance between accuracy and linkage rates. These findings suggest that error rates are sensitive to both data quality and linkage techniques.

## 2. Historical Record Linkage

Record linkage has long been a tool for social science research. Early methods involved first identifying a group of individuals in a manuscript census schedule and then searching, by hand, within the same township, county, or state to locate those individuals in the following census (see Malin 1935, Curti 1959, Bogue 1963, Thernstrom 1964, 1973, and Guest 1987 for examples). The greatest shortcoming of this approach is the inability to link individuals who relocate between census years, which poses a serious threat to the representativeness of the matched sample. Once state-level indexes became available, researchers could create a sample of households from one census and link them backwards in time using birthplaces of children at least 10 years old to determine which state to search within in the previous census (see Steckel 1988 or Schaefer 1985 for examples). This allowed researchers to begin with a more nationally representative sample from the later census year but, like the first technique, geographic mobility

(between birth and the first census year) threatened representativeness. National indexes and Public Use Microdata Samples (PUMS) allowed creation of nationally representative linked samples, such as the linked 1850-1860 sample created by Ferrie (1996).

Ferrie's (1996) approach relies on phonetically coded first and last names, implied year of birth (or age), and state or country of birth to link person records. It also required that if a person was the head of a household in one census year, that they are head of a household in the following census. Variations of this methodology have become ubiquitous with matching in the economic history literature (Stewart 2006; Long 2005, 2006, 2008; Boustan et al. 2012; Abramitzky et al. 2013, 2012; Abramitzky et al. 2014). The more recent uses of this method incorporate text-sting comparators and do not use household composition (e.g., Long and Ferrie 2013). Many researchers use Ancestry.com's search engine (Collins and Wanamaker 2014; Kosack and Ward 2014; Bailey et al. 2011) to conduct record linkage instead, or use MPC's Integrated Public Use Microdata Samples (IPUMS) Linked Representative Samples (Saperstein and Gullickson 2013; Boustan and Collins 2014; Ruggles 2011), both of which employ text-string comparators. Others use their own approach. For instance, Antonie et al.'s (2014) approach uses a text-string comparator and estimates probability scores using truth data and Vector Machine Learning. Mill and Stein (2013) use a method that also employs string comparators and scoring of matches, using an Expectation Maximization (EM) algorithm and maximum likelihood estimation to determine the probability of a true match.

The MPC uses a conservative record linkage approach to create the IPUMS Linked Representative Samples of the 1850-1930 U.S censuses. To maximize representativeness, their approach uses limited linking variables (first name, last name, year of birth, race, and birthplace) and accounts for commonness of names and birthplaces. Their process includes many steps.

First, they preprocess the data. This includes consistency checks for age and standardizing first name text strings (Vick and Huynh 2011). Next, they use the FEBRL software to score age and name similarity within data blocked on sex, race, birthplace, and marital status. Once they have potential links scored, they use the SVM to choose true and false links, incorporating the probability of a link given name and birthplace commonness. The last step eliminates cases with numerous potential links. Ultimately, this approach yields accurately linked samples (Goeken et al. 2011).

Wisselgren et al. (2014) use techniques similar to MPC and find their linkages across the 1890 and 1900 Swedish censuses have low error rates. They also test record linkage strategies such as using standardized name strings, using Swedish-specific patronymic naming conventions to improve the availability of children's surnames, and using household information to create links. Wisselgren et al. (2014) find these strategies increase the number of links and, through comparisons between the linked sample and links to digitized parish registries, they determine these strategies result in highly accurate links. Given increasing use of text-string comparators and probabilistic matching methods in the literature, I begin with a very standard linkage approach and determine the benefits of incorporating these newer techniques.

## 3. Assessing Existing Techniques

### 3.1. Creating Benchmark Linked Data

This analysis uses the 2013 Numident and the 2005 CPS ASEC to create the benchmark linked data. The Numident is a record of nearly 500 million SSNs and contains full name, full date of birth, sex, state or country of birth, and date of death.[4] The Numident also records each

[4] The Census Bureau's authority to obtain the Numident is Title 13, Section 6. Titles 5, 12, and 42 of the U.S. Code give SSA authority to share the Numident with the Census Bureau.

transaction, or claim, on a SSN; therefore, there are records for each instance an individual changes, or corrects, their name or date of birth. Each SSN in the Numident corresponds one-to-one with a Protected Identification Key (PIK) assigned by the Census Bureau. Once assigned a PIK, the Census Bureau removes PII from the record, which protects respondent PII while allowing researchers to use the PIK for record linkage purposes.

The 2005 CPS ASEC consists of 210,648 individuals from 98,664 households. The 2005 CPS ASEC data contain first and last name, sex, age, year of birth, and – for those aged 15 and up – SSN.[5] I create the benchmark sample by linking individuals who provided an SSN in the 2005 CPS ASEC to their Numident record using the PIKs. I use the subset of PIKs from records with verified SSNs to ensure accuracy of the benchmark links. Verification requires that the SSN, name, and date of birth provided on the CPS ASEC file match the information provided in the Numident.[6] Of the 157,804 respondents aged 15 and up in the 2005 CPS ASEC, only 32.3 percent provided a verified SSN necessary for accurate linkage to the Numident.[7]

To create the final sample of benchmark links, I drop observations missing their first name or last name[8] and I calculate implied year of birth as 2005 minus age. The CPS ASEC recorded country of birth, but not state of birth; therefore, I append state of birth from the Numident to the CPS ASEC records using the PIKs to link the CPS ASEC to the Numident. In other contexts, survey-reported birthplace generally agrees with birthplaces in the Numident. For example, place of birth of 95.0 percent of native-born respondents and 83.7 percent of foreign-

[5] The 2005 CPS ASEC is the most recent CPS file that collected SSNs. Beginning in 2006, the CPS stopped collection of SSNs.
[6] For a discussion of the SSN verification process of the Census Bureau, see Wagner and Layne (2014).
[7] SSN was provided by 56,945 (27.0 percent) of respondents in the 2005 CPS ASEC. Of these SSNs, 52,634 (92.4 percent) were verified with the SSA Numident data. The number of verified SSNs may be low because the respondent may not know the SSNs of each member in their household. Table 1 shows that household heads and spouses make up the majority of men in the verified sample, which is consistent with the hypothesis that respondents may only know their own or their spouses SSN.
[8] There was only 1 respondent missing first name, 12 missing last name, and 0 missing age.

born respondents in the Census 2000 Long Form match the place of birth reported in the Numident. Nevertheless, appending birthplace from the Numident to the CPS ASEC may result in slightly higher accuracy rates achieved here than would be expected using respondent-provided place of birth, since place of birth will not be a source of error in the linkage in this analysis.

To emulate Ferrie (1996), I drop all women from the sample. I also address common names by eliminating all observations with identical combinations of first name and last name in the CPS. Ultimately, the CPS ASEC-Numident benchmark data consists of 24,861 men.

Table 1 compares men in the final sample of benchmark links to men in the full 2005 CPS ASEC. The final sample of benchmark links has a larger representation of white, married, and native-born men, whereas non-whites, Hispanics, children, and the foreign-born are underrepresented. Given linkage rates are typically lower for minorities and the foreign born in the literature (Goeken et al. 2011; Abramitzky et al. 2012), the linkage rates for the verified sample may be slightly higher than they would be for the entire CPS ASEC sample.

### 3.2. Methodology

Historians use characteristics that remain unchanged over time (e.g., place of birth) or change predictably (e.g., age) to link individuals across censuses. I use first name, last name, year of birth, and state or country of birth as linking variables in this analysis. To conduct the linkage, I use the standard technique that has evolved from Ferrie (1996), which became the standard in historical record linkage and employed an algorithm similar to the following:

1. Restrict sample to males (also limit age or location depending on research needs).
2. Code names phonetically (either using NYSIIS or SOUNDEX systems).

3. Eliminate common names.[9]
4. Exact match using phonetically coded name and compare the characteristics of potential matches.
    a. Create a band around implied year of birth, drop potential matches falling outside.
    b. Drop potential matches whose birthplaces do not match.
5. If two or more potential matches remain, keep the potential match with the most similar implied year of birth. Drop records that have more than one potential match with identical phonetic codes for first and last name, implied year of birth, and birthplace.

To establish baseline linkage results, I first link CPS ASEC data to the Numident using the approach outlined above, henceforth the "standard approach." Often, year of birth is not available in older data, so researchers use implied year of birth estimated from reported age.[10] I calculate implied year of birth in the CPS ASEC as the survey year (2005) minus age. In the literature, linkages across decennial census data are conducted using age and assuming a ten-year difference in age between two adjacent censuses (Ferrie 1996; Roy 2013). Calculating implied year of birth is essentially the same exercise. The standard approach typically allows the implied year of birth for potential matches (from the exact match on phonetic name) to fall within a chosen band, or interval. I conduct the simulation three times, allowing for one-year, three-year, and five-year intervals.[11] To determine the error rate, I compare the links made using the standard approach to the benchmark links made using SSNs. These simulations are completely automated and do not employ clerical review to determine linkages.

[9] I remove all individuals who are not unique on phonetically coded name, age, and birthplace following Abramitzky et al. (2012). Ferrie (1996) allows no more than 10 identical name combinations, regardless of age or birthplace.
[10] Self-reported year of birth was collected in the 1900 and 1910 U.S. decennial censuses. Age at last birthday is available when year of birth is not.
[11] Ferrie (1996) drops all potential matches with differences between age less than five or greater than fifteen. More recent papers, such as Abramitzky et al. (2012), use one- to five-year bands.

### 3.3 Baseline Linkage Rates

Table 2 reports baseline linkage rates achieved by the standard approach. The standard approach with a one-year band around implied year of birth results in a linkage rate of 41.5 percent of which 81.2 are correct according to the benchmark data. When I increase the band to three years, the standard approach results in a linkage rate of 42.4 percent of which 79.7 percent are correct. The linkage rate increases slightly to 42.9 percent with a five-year band and has an accuracy rate of 79.0 percent.

The accuracy of the baseline sample is much lower than Wisselgren et al. (2014) and Goeken et al. (2011). One potential explanation is that Wisselgren et al. (2014) use a more detailed level of geography for birthplace (parish) in their record linkage across the 1890 and 1900 Swedish Censuses. Sweden had approximately 2,600 parishes at the end of the nineteenth century (Wisselgren et al. 2014, pg. 142) for a population of approximately 5 million people, which is significantly more refined than state or country of birth in the Numident. Goeken et al. (2011) may achieve higher accuracy rates due to matching a June census to another June census. This eliminates much error in the measurement of age, which I show significantly improves accuracy in Section 4.2.

These baseline linkage rates are also higher than what is generally found in the literature, which range anywhere from 3 percent to 39.4 percent across historical censuses.[12] There are several potential reasons why linkage rates for the CPS ASEC are higher. First, the CPS ASEC-Numident links are made with self-reported data. As a result, the name fields may be of higher quality. Also, there is no noise in birthplace in the CPS ASEC because I appended birthplaces

[12] The MPC achieves a match rate of 3 percent for foreign-born males between the 1870 and 1880 censuses (MPC, 2010). Guest (1987) achieves a match rate of 39.4 percent across the 1880 and 1900 censuses. Maloney (2001) achieves a 58 percent match rate between white men living in Cincinnati in the 1920 Census and WWI selective service registration records.

from the Numident. Second, the Numident is a record of all SSNs and the SSA does not remove migrants or the deceased. Therefore, my linkage rates are not subject to loss from attrition. Given that date of death is available in the Numident data, however, I can explore how attrition from death may affect linkage rates across censuses.

### 2.2 Sample Attrition and Linkage Rates

In the literature, explanations for low linkage rates often include migration and death (Ferrie 2004, Guest 1987). I do not observe whether an observation in the CPS ASEC has moved abroad between 2005 and 2013, but I do observe deaths reported to the SSA. This allows me to determine linkage rates were my data subject to attrition from death that would occur between historical censuses.

To test how death affects linkage rates, I eliminate deceased observations from the Numident and rerun the simulation. The Numident provides date of death for a majority of states.[13] If I remove all observations with a date of death between March 1, 2005 and March 1, 2013 from the Numident, and then match it to the 2005 CPS ASEC using the standard method with a one-year band, the linkage rate increases from the original 41.5 percent to 43.9 percent. Only 78.1 percent of these linkages are accurate.

Instead of decreasing the linkage rate, eliminating deceased observations from the 2013 data increased the linkage rate by 2.4 percentage points. This increase in links was accompanied by a decrease in the percentage of correct links by 3.1 percentage points. Upon further examination, all observations alive in both 2005 and 2013 receive the same match they received in the previous specification before removing deceased records. The increase in linked cases

[13] Alaska, Colorado, Maine, Maryland, Massachusetts, Mississippi, New York, North Carolina, Pennsylvania, Rhode Island, Tennessee, Virginia, West Virginia, and Wyoming do not report date of death in the Numident.

results from reducing the pool of potential matches, which allows observations with more common combinations of the linkage variables to find a link. These findings suggest attrition from death is not a significant cause of low linkage rates, and in some cases, may actually increase the linkage rate and the error rate.

## 4. Techniques for Improvement

In addition to attrition, there are multiple reasons why historical linking methods result in low linkage rates. Foremost, the necessity to eliminate common names reduces the linked sample significantly. Also, noise in the data—from misreported age, enumerator error, or keying errors—also adversely affects linkage rate and quality.[14] Despite these issues, several techniques may improve linkage rate and quality.

### 4.1. String Comparators

Dropping common names substantially decreases the number of matches. Without additional information to match on, it is difficult to distinguish between two or more potential links. When additional variables are not available, variation in the spelling of names can provide another means to distinguish between potential linkages.

Jaro-Winkler text-string comparators serve as a measure of how closely two text strings match, while allowing for some degree of misspelling (Winkler 1995). To test the effectiveness of string comparators, I link the CPS ASEC data to the Numident using the same criteria outlined in Section 3.2, but now incorporate a Jaro-Winkler-type text-string comparator. I employ the

[14] See Ewbank (1981), Budd and Guinnane (1991), Stockwell and Wicks (1974), and A'Hearn, Baten, and Crayen (2009) for more information on the effects of misreported age and age heaping on demographic analyses. Age misreporting is particularly high for African Americans (Elo and Preston 1994; Coale and Rives 1973).

string comparator developed by Fiegenbaum (2014) for Stata, requiring a similarity score of at least 9 out of 10 to match the MPC's Jaro-Winkler cutoff value (Goeken et al. 2011).

In this implementation, I first identify all potential matches using phonetic name, implied year of birth, and birthplace. Then, I keep the potential match with the lowest combined distance measure (closest match) for first and last names and the closest implied year of birth. I drop any CPS ASEC observation that has two or more potential matches in the Numident with the same string distance for first and last name. Although it is possible for similar text strings to have phonetic codes that do not match (Mill 2013), I maintain the standard approach's necessity for identical phonetic first and last name. This allows me to isolate the value added by using text-string comparators.

Table 3 reports the linkage results from introducing Jaro-Winkler distances. The linkage rate with a one-year band increases from 41.5 percent in the baseline to 63.0 percent using the string comparator. The percentage of accurate links fell from 81.2 percent to 78.3 percent.

String comparators allow for additional noise in the name fields, which may cause this observed decrease in accuracy. This approach determines the best potential match by how closely the first and last name text strings match. For the majority of cases, this will result in an increase the number of correct linkages. For a small number of cases, this may also introduce some incorrect links, particularly if one of the name fields is misspelled.[15] The increase in error from string comparators may also result from noise in the implied year of birth calculated in the CPS ASEC. I compare implied year of birth, calculated as 2005 minus age in the CPS ASEC, to

[15] Consider two fictitious people, John Smith and Jon Smith (both born in 1955 in Texas). If you observe John Smith in the CPS ASEC and both John Smith and Jon Smith in the Numident, then no match would be found for John Smith in the CPS ASEC using phonetic codes because it is impossible to distinguish between the two potential links. If using string comparators, then John Smith in the CPS ASEC would be linked to John Smith in the Numident. However, if John Smith was misspelled as "Jon Smith" in the CPS ASEC, then this keying error would lead to an erroneous link to Jon Smith in the Numident.

self-reported year of birth in the Numident to conduct the record linkage. As a result, year of birth is measured imprecisely for many of the CPS ASEC observations. Use of string comparators may exaggerate this source of error.

### 4.2. Adjusting Implied Year of Birth

Age changes predictably over time and is useful for record linkage, but age can be collected at different times of the year for different surveys. Therefore, estimating implied year of birth as the survey year minus age introduces additional noise into the linking variables.[16] Depending on the data sources one wants to link, this noise can cause significant error. For example, if I were linking the 1900 census, collected in June, to the 1920 census, collected in January, implied year of birth for these censuses will be noisier than implied year of birth calculated and used to link an April census to another April census. For the 2005 CPS ASEC used here, 2005 minus reported age on March 1, 2005 correctly estimates year of birth for only 17.4 percent of the verified CPS ASEC sample.

*Linking on Age on March 1, 2005*

Using the benchmark data, I determine the extent to which estimated implied year of birth introduces error into the linkages. I calculate age on March 1, 2005 for each observation in the Numident (which contains exact date of birth). This allows a comparison of reported age in the 2005 CPS ASEC to the calculated age on March 1, 2005 in the Numident. In this linkage, I use the standard record linkage approach to link the CPS ASEC to the Numident using phonetic name, place of birth, and age. I report the results of this linkage in Table 4.

[16] Similar inaccuracies arise when others match on age and expect age to be 10 years different between censuses.

Reducing the noise surrounding implied year of birth increases the accuracy of the match significantly. When allowing a one-year interval around age, I match 42.5 percent of the verified sample. Of these, 93.4 percent received the correct match, which is an increase of 12.2 percentage points from the baseline results. When the interval around age is increased to three years, the linkage rate increases to 43.5 percent with 91.6 percent accurate. The linkage rate increases to 44.0 percent, of which 90.6 percent are accurate, with a five-year interval around age. In terms of accuracy, these results may be more representative of the record linkage conducted across two historical censuses taken in the same month. These results also imply the majority of error in the baseline results stems from noise in the implied year of birth variable rather than from the standard approach itself.

*Linking on Age with String Comparators*

Table 5 shows the linkage rates using a string comparator while linking on age. When using string-comparators, the elimination of noise in implied year of birth results in more links, but accuracy is not as high as in Table 4. Using this approach, 66.9 percent of records were linked, with 90.1 percent accurate (allowing a one-year interval around age). When permitting a three-year interval around age, the linkage rate is 67.8 percent with 89.0 percent accurate. The linkage rate increases to 68.2 percent with 88.5 percent accurate with a five-year interval.

*Linking on Adjusted Implied Year of Birth*

Not every data set allows the researcher to adjust age or implied year of birth as accurately as the Numident. For some datasets, there are easy changes that may improve accuracy. If a researcher links a data set with “age at last birthday” to a dataset that asked for

actual year of birth, depending on the time of year the survey collected age, the researcher can adjust estimated year of birth. For example, if linking the 2005 CPS ASEC to the Numident, implied year of birth calculated as 2005 minus age may be wrong for anyone born between March and December, depending on the date of the interview. To decrease the potential for error, the researcher can estimate implied year of birth as 2004 minus age at last birthday, henceforth edited year of birth. With this adjustment, estimated year of birth will now be correct for the majority of observations.

The results from using the standard approach with edited year of birth are reported in Table 6. Allowing a one-year band around edited year of birth resulted in 42.1 percent of observations receiving a match and 91.4 percent of links received the correct match. This represents a significant improvement in accuracy over the standard approach that does not adjust implied year of birth. When I incorporate Jaro-Winkler string comparators into this match, the percent linked increases to 67.1 percent, but the percent accurate decreases to 87.0 percent.

### 4.3. Probabilistic Matching Techniques

Probabilistic matching techniques can improve the linkage rate and accuracy of historical record linkage. In Fellegi and Sunter's (1969) framework, the probabilistic matching algorithm compares two data sources, A and B. Each observation in A is compared to each observation in B, creating a comparison space composed of the Cartesian product of A and B. For each attribute, the algorithm assigns agreement and disagreement weights, determined by the data, according to the similarity of the attribute across observations in the comparison pair. In the case of first and last name, the algorithm assigns an agreement or disagreement weight depending on the Jaro-Winkler distance between the two strings.

Next, the algorithm calculates an overall score for the comparison pair by summing the agreement and disagreement weights over all comparison attributes. If the overall score given to a comparison pair falls above a given cutoff value (chosen by the researcher), that pair is kept as a potential match. At the end of the process, the algorithm keeps only the highest-scoring comparison pair for any given observation in the input data (the CPS ASEC in this analysis). If an input observation has two or more potential matches with identical scores, that observation is dropped from the final sample.

In the application of the probabilistic matching method used here, I process the data through several passes, where each pass limits the comparison space by different combinations of variables. Limiting the size of the comparison space is accomplished by blocking, which increases efficiency and results in higher linkage rates (Michelson and Knoblock 2006).[1] I block by sorting the CPS ASEC and Numident data and breaking them into pieces depending on the characteristics sorted on. Only observations falling into the same blocks, or pieces, are compared in the linkage. Table 7 describes the blocking variables and linkage rates for each pass. I construct five passes, with each pass slightly easing the linkage constraints. Only those records not linked in the first pass proceed to the next, and the records cascade similarly through all five passes.

Of all methods analyzed, the probabilistic algorithm results in the greatest number of linked observations. The probabilistic algorithm matches 15,612 observations, or 62.8 percent, when I match using the edited and implied year of birth. Of these, 93.1 percent receive the correct match. If I rerun the match and substitute age on March 1, 2005 for edited and implied year of birth, the match rate increases to 63.2 percent, of which 94.6 percent are accurate. In

comparison with the other methods assessed in this analysis, the probabilistic approach results in substantially more linked observations without compromising accuracy.

## 5. Discussion of Additional Linking Variables

There are often variables in the data that can enhance linking. For example, many techniques use race or parents' birthplace (Long and Ferrie 2014; Goeken et al. 2011). Although additional variables increase the number and quality of linkages, additional variables may only be available for subsets of the population and their use may introduce bias into the linked sample. Ultimately, researchers may have to compromise representativeness for accuracy and linkage rate if using additional variables.

Several studies link person records using additional linking variables, which have included parents' birthplaces and race. Parents' birthplaces were collected in the 1880-1930 U.S. censuses and the 1940-1960 long form U.S. censuses. Although earlier censuses did not collect parents' birthplace, researchers can reconstruct them for children living with their parents using the parent's birthplace responses. Appending parents' birthplaces to their children would help link them forward to later censuses, but, at the same time, may introduce bias into the linked sample. For example, if you are interested in linking 5-10 year olds in the 1850 census to their adult observations in the 1900 census, appending parents' birthplaces from the household head and spouse would introduce little bias. However, if linking 15-20 year olds in 1850 to 1900, the use of parents' birthplaces could increase the probability of successfully linking individuals who were more likely to be living with their parents in 1850.

Linking on race can also provide additional information to distinguish between potential matches. However, a growing body of literature shows racial identity can be fluid over time

(Liebler et al. 2014) and so the use of race as a linkage key may bias the matching algorithm towards linking individuals with stable racial identities. The Numident does contain some information on race, but race was not collected for those enumerated at birth beginning in 1987 (Scott 1999).[17] I reran the standard approach from Table 2 with the restriction that race in the Numident must match rate in the CPS ASEC and I drop observations under the age of 18.[18] I find accuracy does not change greatly from using race. By blocking on race and allowing a one-year band in implied year of birth, I correctly match 7,549 of 9,198 links (82.1 percent) compared to the 8,015 of 9,682 (82.8 percent) correctly matched by the standard approach without race as a linkage key (also limited to those 18 years and older). Given no observable change in accuracy and the potential to introduce bias, these results suggest researchers may want to reconsider using race as a linkage key.

## 6. Consequences of Incorrect Linkages

Record linkage is useful for creating longitudinal data from cross-sectional data, as well as obtaining additional variables from other datasets. However, if the record linkage process results in inaccurate linkages, use of linked data will introduce measurement error that will bias estimation (Abowd and Vilhuber 2005; Campbell 2008; Kim and Chambers 2012). In this section, I provide an example of how the different record linkage techniques affect the linked samples and their use in regression analysis.

Table 8 provides descriptive statistics of the matched sample from each record linkage technique allowing a one-year difference in implied or edited year of birth. I report descriptive

[17] Race was not collected for anyone enumerated by the SSA at birth beginning in 1987 unless they apply for changes to their SSN later in life (e.g., names changes).

[18] To make the CPS ASEC detailed race codes match those of the Numident, I linked the CPS ASEC to the Numident by SSN and compared the detailed race codes to the race codes in the Numident. I recoded each detailed race in the CSP ASEC to match the race most often associated with that detailed race code in the Numident (by looking at a cross tabulation of detailed race in the CPS ASEC and race in the Numident).

statistics from a mix of variables from the CPS ASEC and those obtained through linkage to the Numident. The Numident contains few variables not measured by the CPS ASEC. One exception is city of birth. To assess the bias introduced by the linkage technique, I create a dummy variable for metropolitan status of city of birth. I also report race as reported in the Numident in addition to several characteristics measured in the CPS ASEC.

There are substantial differences between the benchmark sample and the samples linked by the various techniques. In particular, the percentage Hispanic and the percentage born in metropolitan areas (measured by the Numident) vary widely across the matched samples. Some of these differences arise from the linkage process (e.g., non-minority observations are more likely to be successfully linked), while some results from a CPS ASEC record linked to the incorrect Numident record. For the CPS ASEC variables, there is variation in the percentage of observations that are black alone, Hispanic, have less than a high school diploma, and have a bachelor's degree across the linkage specifications. These results suggest the matched samples are not entirely representative of the benchmark sample.

Measurement error introduced from record linkage biases demographic and economic analyses. In Table 9, I run simple regressions of log total wage and salary earnings (from the CPS ASEC) on race and city-of-birth metropolitan status obtained through linkage to the Numident. I drop observations missing a city of birth and those with unknown or missing race from the Numident. These regressions show substantial variation in the magnitude of the coefficients across matched samples, particularly for the black-white earnings gap and the earnings gap between American Indian Alaska Native men and white men. For the benchmark sample, the simple regression estimates that the earnings of black men are 72.0 percent of the earnings of white men. Across the matched samples, the coefficient suggests black men earn

anywhere between 78.3 percent (probabilistic match) to 86.3 percent (baseline match) of white men. Even the most accurate linkage approach, the probabilistic match, resulted in an attenuated coefficient, suggesting that even a small percent of incorrect linkages introduces bias. The magnitude of these coefficients is robust to weighting the matched sample by the probability of linkage.[19]

## 7. Conclusion

Errors in record linkage introduce measurement error that can bias estimates (Abowd and Vilhuber 2005; Campbell 2008; Kim and Chambers 2012). It is important to identify sources of record linkage error and test possible solutions to limit measurement error. This paper creates benchmark, linked data and assesses the record linkage techniques most commonly used by historians. The benchmark data consist of 2005 CPS ASEC records linked to the Numident by SSN. I establish baseline results by matching on phonetic codes for first and last name, implied year of birth, and place of birth. Then, I incorporate innovations in the record linkage literature, specifically a Jaro-Winkler string comparator and probabilistic matching, to assess their effect on the linkage rate and accuracy.

From a simple match using phonetically coded first and last name, age, and place of birth, I match over 40 percent of men in the benchmark 2005 CPS ASEC sample to the Numident. Comparisons with the benchmark links show approximately 81.2 percent of these linkages are correct. This high amount of error results from measuring age at different points in time in the CPS ASEC and Numident, and I would expect lower error rates if matching from a census collected in April to another census collected in April. I show that if I match age in the 2005 March CPS ASEC to age on March 1, 2005 in Numident, the percentage of accurate links

[19] I follow Long and Ferrie's (2013) approach to construct the weights.

increases to 93.4 percent. This may be a more representative of the error rates from linkage of two April censuses, which is more common in the historical literature and would use age collected at the same time of the year in the linkage. If the data are collected at different times of the year, adjusting implied year of birth (estimated from age) can greatly increase the percentage of accurate links (from 81.2 percent to 90.1 percent for the CPS ASEC-Numident links). The linkage rate increases (from 41.5 percent to 63.0 percent) when I incorporate Jaro-Winkler string comparators; however, accuracy suffers as a result. Using probabilistic matching techniques, I achieve a linkage rate of 62.8 percent with a 93.1 percent accuracy rate.

Although these results inform upon the accuracy of historical linkage techniques, there are several nuances between the techniques and data used here and those used to link historical data to keep in mind. For instance, I use no clerical review to determine the linkages. Since many linkage projects still employ some degree of clerical review, the linkage rates may not be directly comparable. In addition, this analysis uses contemporary data and a select sample of the 2005 CPS ASEC that provided SSNs. As with any linking project, the linkage process is highly reliant on the quality of the information collected in the data, and the conclusions drawn from one dataset may not be extensible to others, particularly if the underlying populations are significantly different. In Wisselgren et al. (2014), for example, they confirm 97.5 percent of men linked across the 1890 and 1900 Swedish censuses with parish records. While none of the methods used here produced accuracy rates as high as 97.5 percent, this difference may result, in part, from linking the CPS ASEC data using state of birth, which is a less fine definition of geography than parish of birth. As a result, my analysis of different linkage techniques should be interpreted with respect to the baseline linkages.

Ultimately, comparisons to the baseline results suggest the standard linkage techniques using phonetically coded name, age, and place of birth achieve the lowest linkage rates. Depending on the quality of the linkage keys, the error from this match can be high. Although newer techniques may increase the number of linked records, I find the cost of this increase can be accuracy conditional on how the match is specified. Overall, I find probabilistic matching techniques provide the best balance of accuracy and high linkage rates.

References

Abowd, J. M. and Vilhuber, L., 2005. The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers. *Journal of Business and Economic Statistics,* 23(2): 133-165.

Abramitzky, R., Boustan, L., Eriksson, K., 2013. Have the Poor Always Been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration. *Journal of Development Economics,* Volume 102: 2-14.

Abramitzky, R., Boustan, L. P. and Eriksson, K., 2012. Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review,* 102(5): 1832-56.

Abramitzky, R., Boustan, L. P. and Eriksson, K., 2014. A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy,* Volume Forthcoming.

A'Hearn, B., Baten, J. and Crayen, D., 2009. Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital. *Journal of Economic History,* 69(03): 783-808.

Antonie, L., Inwood, K., Lizotte, D. J. & Ross, J. A., 2014. Tracking people over time in 19th century Canada for longitudinal analysis. *Mach Learn,* Volume 95, pp. 129-146.

Bailey, A. K., Tolnay, S. E. & Laird, D. J., 2011. Targeting Lynch Victoms: Social Marginality or Status Transgressions?. *American Sociological Review,* 76(3), pp. 412-436.

Bogue, A., 1963. *From Prairie to Corn Belt; Farming on the Illinois and Iowa Prairies in the Nineteenth Century.* Chicago: University of Chicago Press.

Boustan, L., Kahn, M. E. and Rhode, P. W., 2012. Moving to Higher Ground: Migration Response to Natural Disasters in the Early Twentieth Century. *American Economic Review Papers and Proceedings,* 102(3): 238-244.

Boustan, L. P. and Collins, W. J., 2013. The Origins of Persistence of Black-White Differences in Women's Labor Force Participation. *NBER Working Paper No. 19040,* May.

Budd, J. W. and Guinnane, T., 1991. Intentional Age-Misreporting, Age-Heaping, and the 1908 Old Age Pensions Act in Ireland. *Population Studies,* Volume 45: 497-518.

Campbell, K. M., 2009. Impact of record-linkage methodology on performance indicators and multivariate relationships. *Journal of Substance Abuse Treatment,* 36(1): 110-117.

Coale, A. J. and Rives, N. W., 1973. 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,* 39(1): 3-36.

Collins, W. J. and Wanamaker, M. H., 2014. Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data. *American Economic Journal: Applied Economics,* 6(1): 220-252.

Curti, M., 1959. *The Making of an American Community; a Case Study of Democracy in a Frontier County.* Stanford: Stanford University Press.

Elo, I. T. and Preston, S. H., 1994. Estimating African-American Mortality from Inaccurate Data. *Demography,* 31(3): 427-458.

Ewbank, D. C., 1981. *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis,* Washington, D.C.: National Academy Press.

Feigenbaum, J., 2014. JAROWINKLER: Stata module to calculate the JaroWinkler. *Boston College Department of Economics Statistical Software Components S457850*

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

Ferrie, J., 1996. A New Sample of Americans Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules. *Historical Methods,* Volume 34: 141-56.

Ferrie, J. P., 2004. *Longitudinal Data for the Analysis of Mobility in the U.S., 1850-1910.* [Online] Available at: http://faculty.wcas.northwestern.edu/~fe2r/papers/saltlakecity.pdf

Goeken, R., et al. 2011. New Methods of Census Record Linking. *Historical Methods,* 44(1): 7-14.

Guest, A. M., 1987. Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 20(2), pp. 63-77.

Kim, G. and Chambers, R., 2012. Regression Analysis under Probabilistic Multi-Linkage. *Statistica Neerlandica,* 66(1): 64-79.

Kosack, E. and Ward, Z., 2014. Who Crossed the Border? Self-Selection of Mexican Migrants in the Early Twentieth Century. *Journal of Economic History,* 74(4): 1015-1044.

Liebler, C. A. et al., 2014. America's Churning Races: Race and Ethnic Response Changes between Census 2000 and Census 2010. *Forthcoming CARRA Working Paper Series.*

Long, J., 2005. Rural-urban migration and socioeconomic mobility in Victorian Britain. *Journal of Economic History,* 65(1): 1-35.

Long, J., 2006. The socioeconomic return to primary schooling in Victorian England. *Journal of Economic History,* 66(4): 1026-1053.

Long, J., 2008. Social mobility within and across generations in britain since 1851.

Long, J. and Ferrie, J., 2013. Intergenerational Occupational Mobility in Great Britain and the United States since 1850. *American Economic Revie,* 103(4): 1109-37.

Malin, J., 1935. The Turnover of Farm Population in Kansas. *Kansas Historical,* Volume 20: 339-372.

Maloney, T. N., 2001. Migration and Economic Opportunity in the 1910s: New Evidence on African-American Occupational Mobility in the North. *Explorations in Economic History,* 38(1), pp. 147-165.

Michelson, M. and Knoblock, C. A., 2006. Learning Blocking Schemes for Record Linkage. *Proceedings of the 21st National Conference on Artificial Intelligence,* Volume AAAI-06.

Mill, R., 2013. Record Linkage across Historical Datasets. In: *Inequality and Discrimination in Historical and Modern Labor Markets.* Stanford: Thesis (Ph.D.) Stanford University, p. 114.

Mill, R., and Stein, L. C., 2013. Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America. In: *Inequality and Discrimination in Historical and Modern Labor Markets.* Stanford: Thesis (Ph.D.) Stanford University, p. 7.

Minnesota Population Center, 2010. *IPUMS Linked Representative Samples, 1850-1930.* [Online] Available at: https://usa.ipums.org/usa/linked_data_samples.shtml
[Accessed 10 2013].

Rastogi, S. and O'Hara, A., 2012. *2010 Census Match Study,* Washington, DC: United States Department of Commerce.

Ruggles, S., 2006. Linking historical censuses: A new approach. *History and Computing,* Volume 14: 213-24.

Ruggles, S., 2011. Intergenerational coresidence and family transitions in the united states,. *Journal of Marriage and Family,* 73(1), p. :136–148.

Saperstein, A. and Gullickson, A., 2013. A Mulatto Escape Hatch? Examining. *Demography,* Volume 50: 1921-1942.

Schaefer, D., 1985. A Statistical Profile of Frontier and New South Migration: 1850-1860. *Agricultural History,* Volume 59: 563-567.

Scott, C. G., 1999. Identifying the Race or Ethnicity of SSI Recipients. *Social Security Bulletin,* 62(4), pp. 1-12.

Steckel, R., 1988. Census Matching and Migration: A Research Strategy. *Historical Methods,* Volume 21: 52-60.

Stewart, J. I., 2006. Migration to the agricultural frontier and wealth accumulation, 1860–1870. *Explorations in Economic History,* 43(4): 547-577.

Stockwell, E. G. and Wicks, J. W., 1974. Age heaping in recent national censuses. *Biodemography and Social Biology,* 21(2): 163-167.

Thernstrom, S., 1964. *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Cambridge: Harvard University Press.

Thernstrom, S., 1973. *The Other Bostonians; Poverty and Progress in the American Metropolis, 1880-1970.* Cambridge: Harvard University Press.

Vick, R. and Huynh, L., 2011. The Effects of Name Standardization on Historical Record Linkage. *Historical Methods,* Volume 44: 15-24.

Wagner, D. and Layne, M., 2014. The Person Identification Validation System: Applying the Center for Administrative Records and Research and Applications' Record Linkage Software. *Center for Administrative Records Research and Applications Report Series (#2014-01).*

Winkler, W. E., 1995. Matching and Record Linkage. In: *Business Survey Methods.* New York: J. Wiley: 355-384.

Wisselgren, M. J., Edvinsson, S., Berggren, M. and Larsson, M., 2014. Testing Methods of Record Linkage on Swedish Censuses. *Historical Methods,* 47(3): 138-151.

| Table 1: Characteristics of the Sample of Benchmark Links | | | | |
|---|---|---|---|---|
| | Verified Benchmark Sample | | All Men in CPS ASEC 15+ Years Old | |
| | N | % | N | % |
| Race | | | | |
| White Only | 21,598 | 86.9% | 61,298 | 81.6% |
| Black Only | 1,679 | 6.8% | 7,533 | 10.0% |
| American Indian and Alaska Native Only | 322 | 1.3% | 1,007 | 1.3% |
| Asian Only | 656 | 2.6% | 3,294 | 4.4% |
| Other Race* | 606 | 2.4% | 2,027 | 2.7% |
| Hispanic or Latino Origin | | | | |
| Hispanic | 1,986 | 8.0% | 10,894 | 14.5% |
| Non Hispanic | 22,875 | 92.0% | 64,265 | 85.5% |
| Family Relationship | | | | |
| Head of Household | 11,921 | 48.0% | 30,463 | 40.5% |
| Spouse | 4,780 | 19.2% | 15,802 | 21.0% |
| Child | 2,558 | 10.3% | 12,528 | 16.7% |
| Other Relative | 377 | 1.5% | 2,623 | 3.5% |
| Not a Family Member | 5,225 | 21.0% | 13,743 | 18.3% |
| Marital Status | | | | |
| Married | 15,584 | 62.7% | 43,492 | 57.9% |
| Widowed | 704 | 2.8% | 1,645 | 2.2% |
| Divorced or Separated | 2,660 | 10.7% | 6,854 | 9.1% |
| Never Married | 5,913 | 23.9% | 23,168 | 30.8% |
| Birthplace | | | | |
| Foreign Born | 22,899 | 92.1% | 11,397 | 15.2% |
| Native Born | 1,962 | 7.9% | 63,762 | 84.8% |
| Total | 24,861 | 100.0% | 75,159 | 100.0% |

Source: Unweighted 2005 CPS ASEC

* “Other” race responses include Hawaiian Pacific Islander Only, some other race, and two or more races.

| Table 2: Baseline Results using the Standard Approach | | | | |
|---|---|---|---|---|
| | Matched | Linkage Rate | Correct | Percent Correct |
| One-Year Band | 10,324 | 41.5% | 8,386 | 81.2% |
| Three-Year Band | 10,551 | 42.4% | 8,413 | 79.7% |
| Five-Year Band | 10,653 | 42.9% | 8,417 | 79.0% |

Source: 2005 CPS ASEC linked to the Numident

Notes: This linked sample was created by linking records in the 2005 CPS ASEC to the Numident using phonetically coded first and last name, implied year of birth (2005-age), and birthplace.

| Table 3: Incorporating Jaro-Winkler String Comparator | | | | |
|---|---|---|---|---|
| | Matched | Linkage Rate | Correct | Percent Correct |
| One-Year Band | 15,662 | 63.0% | 12,270 | 78.3% |
| Three-Year Band | 15,876 | 63.9% | 12,309 | 77.5% |
| Five-Year Band | 15,986 | 64.3% | 12,316 | 77.0% |

Source: 2005 CPS ASEC linked to the Numident

Notes: This linked sample was created by linking records in the 2005 CPS ASEC to the Numident using phonetically coded first and last name, string distance between first and last name, implied year of birth (2005-age), and birthplace. I calculated Jaro-Winkler distances using the Stata program created by Fiegenbaum (2014).

| Table 4: Linking on Age on March 1, 2005 | | | | |
|---|---|---|---|---|
| | Matched | Linkage Rate | Correct | Percent Correct |
| One-Year Band | 10,568 | 42.5% | 9,872 | 93.4% |
| Three-Year Band | 10,808 | 43.5% | 9,903 | 91.6% |
| Five-Year Band | 10,934 | 44.0% | 9,908 | 90.6% |

Source: 2005 CPS ASEC linked to the Numident

Notes: This linked sample was created by linking records in the 2005 CPS ASEC to the Numident using phonetically coded first and last name, implied year of birth (2005-age), and birthplace.

| Table 5: Linking on Age on March 1, 2005 with a String Comparator | | | | |
|---|---|---|---|---|
| | Matched | Linkage Rate | Correct | Percent Correct |
| One-Year Band | 16,635 | 66.9% | 14,983 | 90.1% |
| Three-Year Band | 16,858 | 67.8% | 15,012 | 89.0% |
| Five-Year Band | 16,967 | 68.2% | 15,019 | 88.5% |

Source: 2005 CPS ASEC linked to the Numident

Notes: This linked sample was created by linking records in the 2005 CPS ASEC to the Numident using phonetically coded first and last name, string distance between first and last name, age, and birthplace. I calculated Jaro-Winkler distances using the Stata program created by Fiegenbaum (2014).

| Table 6: Linkage using Edited Year of Birth (2004 – age) | | | | |
|---|---|---|---|---|
| | Standard Approach | | | |
| | Matched | Linkage Rate | Correct | Percent Correct |
| One-Year Band | 10,457 | 42.1% | 9,556 | 91.4% |
| Three-Year Band | 10,720 | 43.1% | 9,589 | 89.4% |
| Five-Year Band | 10,850 | 43.6% | 9,594 | 88.4% |
| | Jaro-Winkler String Comparator | | | |
| | Matched | Linkage rate | Correct | Percent Correct |
| One-Year Band | 16,672 | 67.1% | 14,504 | 87.0% |
| Three-Year Band | 16,676 | 67.1% | 14,505 | 87.0% |
| Five-Year Band | 16,789 | 67.5% | 14,512 | 86.4% |

Source: 2005 CPS ASEC linked to the Numident

| Table 7: Probabilistic Matching Approach | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Match with Edited and Implied YOB | | | Match using Age on March 1, 2005 | | |
| Pass | Blocking Variables | Matched | Correctly Matched | % Correct | Matched | Correctly Matched | % Correct |
| 1 | · First Name<br>· Last Name<br>· Birthplace<br>· Edited YOB | 12,240 | 11,517 | 94.1% | 14,117 | 13,512 | 95.7% |
| 2 | · First Name<br>· Last Name<br>· Birthplace | 2,526 | 2,286 | 90.5% | 682 | 511 | 74.9% |
| 3 | · Truncated First Name<br>· Truncated Last Name<br>· Birthplace<br>· Edited YOB | 380 | 354 | 93.2% | 445 | 419 | 94.2% |
| 4 | · Truncated First Name<br>· Truncated Last Name<br>· Birthplace<br>· Edited YOB | 270 | 234 | 86.7% | 289 | 262 | 90.7% |
| 5 | · Truncated First Name<br>· Truncated Last Name<br>· Birthplace<br>· Implied YOB | 196 | 148 | 75.5% | 170 | 152 | 89.4% |
| Total | | 15,612 | 14,539 | 93.1% | 15,703 | 14,856 | 94.6% |

Source: 2005 CPS ASEC linked to the Numident

| Table 8: Descriptive Statistics of the Matched Samples | | | | | |
|---|---|---|---|---|---|
| | Benchmark Sample | Baseline Sample | Jaro-Winkler Comparator | Standard Approach with Edited Age | Probabilistic Match |
| Total Observations | 24,861 | 10,324 | 15,662 | 10,457 | 15,612 |
| Match Rate | 100.0% | 41.5% | 63.0% | 42.5% | 62.8% |
| Accurate Links | 100.0% | 81.2% | 78.3% | 91.4% | 93.1% |
| | | | | | |
| Numident Variables | | | | | |
| Race | | | | | |
| Unknown | 3.5% | 5.5% | 4.4% | 5.4% | 4.3% |
| White | 79.3% | 81.8% | 80.1% | 82.4% | 81.6% |
| Black or African Origin | 6.7% | 5.1% | 6.9% | 4.8% | 6.4% |
| Other | 1.4% | 1.5% | 1.4% | 1.5% | 1.4% |
| Asian or Pacific Islander | 2.4% | 1.7% | 1.7% | 1.7% | 1.6% |
| Hispanic | 5.5% | 3.2% | 3.9% | 3.0% | 3.3% |
| North American Indian or Alaska Native | 0.8% | 0.6% | 0.8% | 0.6% | 0.7% |
| Missing | 0.5% | 0.7% | 0.8% | 0.6% | 0.6% |
| Place of Birth Metro Status | | | | | |
| Metro | 45.3% | 34.5% | 43.7% | 35.3% | 44.2% |
| Non Metro | 41.1% | 54.6% | 46.2% | 54.0% | 46.5% |
| Missing | 13.5% | 10.9% | 10.1% | 10.7% | 9.4% |
| | | | | | |
| CPS ASEC Variables | | | | | |
| Race | | | | | |
| White Alone | 86.9% | 89.2% | 87.8% | 89.5% | 88.0% |
| Black Alone | 6.8% | 5.2% | 6.6% | 5.1% | 6.4% |
| American Indian Alaska Native Alone | 1.3% | 1.2% | 1.3% | 1.1% | 1.3% |
| Asian Alone | 2.6% | 2.1% | 1.9% | 2.0% | 1.8% |
| Other Race | 2.4% | 2.3% | 2.5% | 2.2% | 2.4% |
| Hispanic or Latino Origin | | | | | |
| Non Hispanic | 92.0% | 94.7% | 93.8% | 94.9% | 94.6% |
| Hispanic | 8.0% | 5.3% | 6.2% | 5.1% | 5.4% |
| Average Age | 45.4 | 44.4 | 45.6 | 44.3 | 45.7 |
| Average Wage and Salary Earnings | 32,749 | 33,389 | 32,335 | 34,021 | 32,588 |
| Education Attainment | | | | | |
| Less than High School Graduate | 16.7% | 14.9% | 16.1% | 14.8% | 15.6% |
| High School Graduate | 29.9% | 29.5% | 30.0% | 29.1% | 30.3% |
| Some College or Associates | 27.2% | 27.9% | 27.6% | 27.9% | 27.8% |
| Bachelor's Degree | 16.8% | 18.1% | 17.0% | 18.5% | 17.1% |
| Advanced Degree | 9.4% | 9.6% | 9.3% | 9.6% | 9.2% |

Source: 2005 CPS ASEC linked to the Numident

Notes: Each matched sample allowed a 1-year interval around implied or edited year of birth.

| Table 9: Wage Regressions using the Matched Samples | | | | | |
|---|---|---|---|---|---|
| | Benchmark Sample | Baseline Sample | Jaro-Winkler Comparator | Standard Approach with Edited Age | Probabilistic Match |
| City of Birth Metropolitan Status | 0.009 | -0.038 | 0.032 | -0.008 | 0.033 |
| | (0.019) | (0.030) | (0.024) | (0.030) | (0.024) |
| Black or African Origin | -0.328*** | -0.147** | -0.169*** | -0.148** | -0.244*** |
| | (0.041) | (0.069) | (0.049) | (0.072) | (0.051) |
| Other | 0.030 | -0.059 | -0.002 | 0.031 | 0.009 |
| | (0.094) | (0.136) | (0.108) | (0.137) | (0.109) |
| Asian or Pacific Islander | -0.553*** | -0.545*** | -0.561*** | -0.520*** | -0.631*** |
| | (0.106) | (0.171) | (0.125) | (0.174) | (0.123) |
| Hispanic | -0.460*** | -0.442*** | -0.355*** | -0.321*** | -0.340*** |
| | (0.054) | (0.105) | (0.072) | (0.107) | (0.073) |
| North American Indian or Alaska Native | -0.789*** | -1.146*** | -1.040*** | -1.104*** | -1.106*** |
| | (0.113) | (0.201) | (0.141) | (0.207) | (0.143) |
| Constant | 10.303*** | 10.326*** | 10.290*** | 10.309*** | 10.297*** |
| | (0.014) | (0.024) | (0.018) | (0.023) | (0.017) |
| Observations | 14,467 | 6,294 | 9,340 | 6,432 | 9,442 |
| R-Squared | 0.014 | 0.010 | 0.011 | 0.008 | 0.013 |

Source: 2005 CPS ASEC linked to the Numident

Notes: The dependent variable is the natural log of total wage and salary earnings. Metropolitan status and race are measured from the Numident. The omitted categories are non-metropolitan city of birth and white. Standard Errors in parentheses: *** p<0.01, **p<0.05, *p>0.1