CARRA Working Paper Series

Working Paper 2014-12

Creating Linked Historical Data: An Assessment of the Census Bureau's Ability to Assign Protected Identification Keys to the 1960 Census

Catherine G. Massey
Center for Administrative Records Research & Applications
U.S. Census Bureau
Catherine.G.Massey@Census.gov

September 2014

Disclaimer: This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Creating Linked Historical Data: An Assessment of the Census Bureau's Ability to Assign Protected Identification Keys to the 1960 Census

Catherine Massey
Center for Administrative Records Research and Applications
U.S. Census Bureau
Catherine.G.Massey@Census.gov

September 2014

Abstract

In order to study social phenomena over the course of the 20th century, the Census Bureau is investigating the feasibility of digitizing historical census records and linking them to contemporary data. However, historical censuses have limited personally identifiable information available to match on. In this paper, I discuss the problems associated with matching older censuses to contemporary data files, and I describe the matching process used to match a small sample of the 1960 census to the Social Security Administration Numeric Identification System.

Keywords: Administrative Records, Record Linkage, Census Samples, Historical Demography

Contents

1. Introduction

Longitudinal data allow researchers to observe individuals at two or more points in time, facilitating social science research. Collections of longitudinal data have increased in recent decades; however, person-level longitudinal data are scarce in historical periods and often lack the observations necessary for precision and the study of smaller populations. The solution is to create linked samples across existing data sets using personally identifiable information (PII).

Historians have long employed matching methods to link individuals across released historical censuses. However, the 1940 census is the most recently released census available to researchers without access to restricted data. For those who do have restricted access to census data, only the 2000 and 2010 censuses contain full names, which are required for matching. While linked contemporary census data are certainly useful for researchers, they are unable to shed light on social phenomena over the course of the 20th century. This paper is the first step in determining the feasibility of transcribing the 1960 census for the purposes of linking individuals across multiple data sets and providing complete microdata files for researchers at the Research Data Centers (RDC).

Transcribing name and partial date of birth from the 1960 short form census would allow researchers to link individuals across several decennial censuses. Those interested in historical analyses could link respondents to their census responses backwards to the released historical censuses, or forward to the 2000 census, 2010 census, American Community Survey (ACS), Current Population Survey (CPS), or the Survey of Income and Program Participation (SIPP).[1] The full 1960 short-form would also provide valuable relationship data needed to create parent-child linkages, thus enabling research on intergenerational mobility.

In addition to facilitating linkage, transcribing the 1960 census would provide an additional year of microdata available to researchers in the RDC. Currently, restricted-use versions of the 1970-2010 microdata are available through the RDC and they provide demographic, household, and detailed geographic information for all individuals surveyed. Although a sample of the 1960 long-form is available for research, the complete census would provide substantially more observations with more precise geographic information (Ruggles, Schroeder, Rivers, Alexander, and Gardner, 2011).[2]

The Census Bureau is uniquely equipped to provide high quality matches across the 1960 census and other data sources. The Center for Administrative Records Research and Applications (CARRA) has developed state-of-the-art person matching software, known as the Personal Identification Validation System (PVS), to facilitate record linkage. PVS uses a probabilistic matching algorithm to link person records in decennial, survey, state, and administrative data to a reference file and append unique Protected Identification Keys (PIK) (Wagner and Layne, 2014). Each PIK corresponds to an SSN in the PVS reference file, which is largely composed of data from the Social Security Administration (SSA) Numerical Identification System (referred to as the Numident). Person records processed by PVS are linkable by PIK to other data sources.

In this paper, I create an altered version of PVS to assign PIKs to a transcribed sample of the 1960 census and assess the PIK rate. First, I modify the PVS software to reflect the limited

[1] The 1850-1940 censuses have been fully released to the public. The Minnesota Population Center has created public use, linked samples across these censuses (see Goeken, Huyhn, Lynch, and Vick, 2011).
[2] The 1960 long form census is the largest long form sample and includes one in four households. Only a one percent sample is currently available for research.

amount of PII available in the 1960 census and incorporate matching techniques employed in the history literature. I consider two treatments for the data: one procedure for the short form census and another for the long form.

Despite the limited amount of information available to match on in the 1960 census, my findings suggest that PIK rate and accuracy would be high for the entire 1960 census.[3] Depending on the variables matched on, I achieve PIK rates ranging from 70.5 percent (for the short form) to 75.5 percent (for the long form). In a simulation using the 2010 Census and the modified PVS, I achieve a PIK rate of 63.6 percent. Of these, 92.8 percent received the same PIK assigned by the Census Bureau's formal PVS.

2. Matching Techniques in History Literature

The majority of matching conducted by CARRA uses full name, full date of birth, addresses, and social security numbers (SSN) to identify and validate potential linkages with the Numident. The combination of address with name and birthdate results in high quality matches and high match rates (Layne, Wagner, and Rothhaas, 2014). However, the 1960 census did not collect SSNs, and no federal administrative data with residential addresses from 1960 are available for the PVS reference file. Matching performed by historians suffers from the same problem of limited PII. As a result, the match rates and techniques used in the history literature can inform upon the necessary modifications to PVS, as well serve as a benchmark to gauge how well PVS performs the match.

Historians typically match person records across censuses using phonetic codes for first and last name, year of birth, place of birth, and, occasionally, other characteristics that are assumed to be time invariant. The majority of the recent literature employs the matching techniques standardized by Ferrie (1996). Ferrie's historical linkage techniques employ the following algorithm:

1) Restrict sample to males (limit age or location depending on research needs)
2) Code names phonetically and truncate name to the fourth letter (either using NYSIIS or SOUNDEX)
3) Eliminate common names
4) Compare characteristics of potential matches
    a. Create a band around estimated (or implied) year of birth, drop potential matches falling outside (varies from 1 to 15 years in the literature)
    b. Drop potential matches whose birthplaces do not match
    c. Drop potential matches based off of other observable characteristics (e.g., if a man is head of a household in an earlier census, he should be head of that same household in the later census)
5) If two or more potential matches remain, keep the match with the closest birth year (described as the "iterative process" in (Abramitzky, Boustan, Eriksson, 2012, 2014))

Table 1 shows a list of techniques used in recent papers. Linkage rates range from 3 to 29 percent in the literature.

[3] I define the PIK rate as the number of observations assigned a PIK divided by the total sample.

Table 1: Matching Techniques used in Recent History Literature

Authors	Year	Data Sources	Matching Technique	Match Criteria	Match Rates
Abramitzky, Boustan, Eriksson	2014	• 1900 US Census • 1910 US Census • 1920 US census	Iterative Approach	• phonetic name • year of birth • country of birth	Match 19% of all native born, 10% of foreign born
Collins and Wanamaker	2014	• 1910 US Census • 1930 US Census	Iterative Approach	• phonetic name • age • birthplace	21%
Long and Ferrie	2013	• 1851 and 1881 UK Census • 1850 and 1880 US Censuses	Variation of Ferrie (1996)	• phonetic name • SPEDIS Scores • age • birthplace	20% in UK, 22% of US
Abramitzky, Boustan, Eriksson	2012	• Norway's 1865 and 1900 Census, • 1900 US Census	Iterative Approach	• phonetic name • year of birth	29% match rate
Minnesota Population Center	2010	• 1850-1930 US Censuses	Freely Extensible Biomedical Record Linkage (FEBRL)	• name • age • birthplace • race	12.2% of white males, 3% of foreign-born males, 6.4% of black males for the 1870-1880 linkage

Some recent papers use the search engine provided by Ancestry.com to conduct their matches (for example, Collins and Wanamaker, 2014). First, they begin with a sample of men from an earlier census and search for these men using Ancestry.com. Ancestry.com allows the researcher to search on name, birthplace, and year of birth with either a one- or two-year band. Once a potential match is located, the researcher loads the scanned image of the enumeration form and transcribes it.

The Minnesota Population Center (MPC) used the Freely Extensible Biomedical Record Linkage (FEBRL) software to link person records across the 1850-1930 censuses (Ruggles, 2006 and 2011). The key innovations of FERBRL include the use of a comparison routine to score the similarity of two text strings (names) and the scoring of potential matches. Unlike phonetic codes, this approach allows the researcher to take advantage of the variation in names that is lost by using phonetic names, and produces higher quality matches (Jaro, 1989; Winkler, 2006).

Long and Ferrie (2013) also employ string comparators in their linkage process. First, they use SOUNDEX codes to sort person records into groups by phonetic code. Then they use string distances calculated from the SAS built-in string comparator, SPEDIS, to examine names within each group. Names scored below a threshold were dropped. The Census Bureau’s matching process similarly uses string comparators and scoring schemes.

3. Person Identification Validation System

3.1. Brief Overview of PVS

CARRA uses the PVS to assign unique PIKs to person records to facilitate unduplication and record linkage. The PVS uses a probabilistic matching algorithm (Felligi and Sunter, 1969) that processes data using the following modules – Verification,[4] GeoSearch, Movers, Name, Date of Birth (DOB), and Household Composition. These modules match data to a reference file based on the Social Security Administration (SSA) Numident file (Wagner and Layne, 2014), using different combinations of personally identifiable information (PII) such as social security numbers (SSNs), name, date of birth, and address. Data cascades through the modules, and only those observations that did not receive a PIK pass from one module to the next. All reference file records are available for linkage in each module. I modify the Name Search module to PIK the 1960 data. The Name Search module uses first name, sex, and date of birth to link records.

3.2. Differences between PVS and Techniques used in History Literature

Several key differences exist between the matching process of PVS and the standard techniques used by historians. First, to compare names across the input and reference file, PVS employs a string comparator program to measure Jaro-Winkler distances between names in the input and reference files (Winkler, 1995). These distances serve as a measure of how closely two names match, while allowing for some degree of misspelling.

There are clear advantages to using a string comparator. Instead of using phonetic codes to match names across two data sources, Jaro-Winkler distances allow the researcher to determine the cutoff value of an acceptable distance between two strings. Furthermore, where "Katheryn," "Catherine" and "Katherine" share the same phonetic code, string comparators allow researchers to take advantage of variation in the spelling of similar names to differentiate between two possible matches.

Second, PVS scores potential matches based on how closely two person records match. Each matching variable has an agreement and disagreement weight.[5] Depending on the similarity between a characteristic of an input observation and reference observation, PVS assigns the agreement or disagreement weight to that characteristic for each comparison pair. PVS gives penalties, or negative scores, when variables in the input and reference file do not match.

PVS calculates a total score for a potential link by summing the agreement and disagreement weights over all variables used in the match. If this score falls above the chosen cutoff value, the PVS software appends the PIK associated with the reference observation to the input observation. If there are multiple potential matches with an overall score that falls above the cutoff, PVS keeps the match with the highest score. PVS does not assign a PIK to input observations with two or more potential matches with the same score, as there is no way to distinguish between the potential matches.

[4] The Verification module is used when data contain reported SSN. An exact match to the Numident is performed by SSN, then the name and DOB are compared. If the name and DOB agree sufficiently, the input record is flagged as verified and the records do not cascade to any other search module. Note that older censuses did not collect SSN, so none of the decennial records were processed through this module.

[5] Agreement and disagreement weights are either dictated by the researcher, derived from machine learning, or estimated by maximum likelihood methods. In this application, weights were determined by clerical review of the data.

Typically, in historical matching, a potential match is deemed a true match if phonetic code, birthplace, and implied year of birth are the same (often with a band around implied year of birth). Because the PVS modules assign a score based on the similarity between characteristics in the input and reference file, the researcher can assign different weights to different characteristics and then test the sensitivity of their choices.[6] The ability for the researcher to test the sensitivity of each decision made during the matching process ultimately decreases the subjectivity of the matching process.

Another difference between PVS and historic matching methods involves treatment of observations with common names. Instead of throwing out all observations of individuals with common names, probabilistic matching techniques in PVS allow researchers to determine the best match using scores that measure the similarity between multiple potential matches for the same entity.

Lastly, PVS employs a high-quality data source as the reference file. The reference file is a database of all SSNs ever issued, with transactions for name and date of birth changes. The reference file includes deceased persons and emigrants; therefore, PIKs are less susceptible to attrition issues caused by these exits. I discuss the reference data and 1960 census data in more detail in the next section.

4. Data

4.1. 1960 Census

This analysis uses two datasets: a sample of the 1960 census drawn from North Dakota and the Social Security Administration's Numident file. The North Dakota 1960 census sample contains 1,727 transcribed entries from the 1960 short form census. I limit the sample to the 1,440 individuals who were also included in the long form. This allows me to compare the match rate achievable for the short form versus the long form, which contains additional information available for matching.[7]

4.2. SSA Numident

The SSA maintains a database of all SSNs ever assigned in the Numident. The Numident contains full name, full date of birth, sex, race, parent's first and last names, and state or country of birth.[8] The Numident also documents when individuals update their information with the SSA. As a result, transactions such as name changes are recorded. Therefore, using the Numident as a reference file is the only data source that allows researchers to accurately link observations of women who change their names when they marry.

[6] For instance, researchers may want to assign more weight to surnames versus middle names, which may result in higher quality matches than applying the same weight to both.

[7] See the appendix for more information regarding creation of the sample used in the analysis.

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

5. Methodology

First, I conduct the match limiting the linking variables to those available in the 1960 short form: first name, middle initial, last name, sex, quarter of birth, and year of birth. I construct a matching process with four passes, where each pass uses different blocking variables.[9]

Blocking reduces the computational burden of linking records (Michelson and Knoblock, 2006). PVS compares each observation in the input data to each observation in the reference data – creating a Cartesian product of the input observations and reference observations. Without breaking up the data, it is computationally difficult to process the total number of potential matches in the Cartesian product. Therefore, PVS sorts the input and reference data by the blocking variables specified for each pass, and only compares observations in the input data to reference data sorted into the same block. As a result, blocking greatly reduces the number of comparisons made.

Table 2 describes the blocking variables used in each pass. In the first pass, I block on NYSIIS first and last name codes, first name, last name, year of birth, and quarter of birth. This is the most restrictive pass of the matching process, and PVS sends individuals who do not receive a potential match in the first pass to the second pass. In the second pass, I block on NYSIIS first and last name codes, last name, and year of birth. The third pass blocks on NYSIIS codes for last name and first name, as well as first and last name text strings. The final pass blocks only on NYSIIS code for first and last name.

Table 2: Blocking Variables

Pass	Blocking Variables	Matching Variables
1	· NYSIIS First Name · NYSIIS Last Name · First Name · Last Name · Year of Birth · Quarter of Birth	· First Name · Last Name · Middle Initial · Year of Birth · Quarter of Birth
2	· NYSIIS First Name · NYSIIS Last Name · Last Name · Year of Birth	· First Name · Last Name · Middle Initial · Year of Birth · Quarter of Birth
3	· NYSIIS First Name · NYSIIS Last Name · First Name · Last Name	· First Name · Last Name · Middle Initial · Year of Birth · Quarter of Birth
4	· NYSIIS First Name · NYSIIS Last Name	· First Name · Last Name · Middle Initial · Year of Birth · Quarter of Birth

[9] Blocking here refers to sorting the input and reference data by different characteristics into blocks. Input observations are only compared to reference observations within the same block.

I block on NYSIIS code in each block to mimic the matching process used in the history literature, which primarily matches using phonetic codes. Even when NYSIIS codes are the same, text strings for names can be substantially different (Mill, 2012); therefore, I also block on last name and first name in several of the passes. I keep the scoring scheme constant across all passes.

Next, I rerun the match and incorporate birthplace as a matching variable. Birthplace is available in the long form census. I use state of birth Federal Information Processing Standard (FIPS) codes and a categorical variable indicating whether an individual was born outside of the U.S. as additional matching criteria. I maintain the same passes and blocking variables as before, with the addition of state of birth and the foreign indicator as additional scoring variables.[10]

After PIKing the 1960 sample, I rerun the match on the subset of the North Dakota 1960 census sample that indicated they were the "child" of the household head. Since the Numident contains parents' first and last names, additional linking variables are available for these children. First, I limit the sample to those indicating they are the child of the household head. To get a baseline match rate for children, I match the subset of children using first name, middle initial, last name, sex, quarter of birth, and year of birth. Next, I include the first and last name of the "household head" and "wife of household head" as potential parent names. I then rerun the match on the sample of children using mothers' and fathers' first and last names as linking variables.[11] Lastly, I conduct the above match on the subset of children using state of birth and the foreign indicator as matching variables.

6. PIK Results

6.1 Full-Sample PIK Rates

Table 3 reports the percentage of observations assigned a PIK by demographic characteristic. When matching on first name, middle initial, last name, sex, quarter of birth and year of birth, 70.5 percent of observations receive a PIK. The matching process does not appear to systematically favor one sex over the other, but some clear differences in PIK rates arise amongst the other observable characteristics. In particular, the PIK rate is low for "other relatives" (50.0 percent), those born in the third quarter (66.2 percent), widowed individuals (54.4 percent), and the foreign-born (63.6 percent).

In the final columns of Table 3, I report PIK rates achieved when including state of birth and the foreign birthplace indicator as additional linking variables. The overall PIK rate increases to 75.5 percent when including birthplace but remains low for other relatives (50.0 percent), those born in the third quarter (71.3 percent), widowed individuals (57.9 percent), and the foreign-born (67.7 percent). The addition of birthplace information improves the PIK rate of wives of

[10] This approach may bias the matching process towards linking individuals born in the U.S.; however, only 6.8 percent of the sample is foreign-born. Furthermore, because birthplace is not available for all individuals in the Numident, if there are characteristics that make an individual more or less likely to have birthplace reported in the Numident, using birthplace would introduce bias into the linkage.

[11] Using household head and wife of household head as parents' names may bias the sample towards linking children from traditional, more stable households because the decennial censuses do not distinguish between household head, wife of household head, and actual parent names.

household heads and those with separated marital status the greatest, but had no effect on the PIK rates for other relatives and non-relatives.

Table 3: Match Rates of Full Sample by Demographic Characteristics

	Full Sample		Matched Without Birthplace		Matched With Birthplace	
	N	%	N	%	N	%
Sex						
Male	732	100%	516	70.49%	559	76.37%
Female	704	100%	496	70.45%	525	74.57%
Missing	4		3		3	
Relationship						
Head of HH	370	100%	249	67.30%	268	72.43%
Wife of Head	293	100%	212	72.35%	231	78.84%
Child	719	100%	519	72.18%	553	76.91%
Other Relative	36	100%	18	50.00%	18	50.00%
Non-Relative	20	100%	16	80.00%	16	80.00%
Missing	2		1		1	
Quarter of Birth						
Jan-March	384	100%	275	71.61%	295	76.82%
April-June	346	100%	251	72.54%	265	76.59%
July-Sept	390	100%	258	66.15%	278	71.28%
Oct-Dec	320	100%	231	72.19%	249	77.81%
Race						
White	1,429	100%	1,008	70.54%	1,080	75.58%
Black	1	100%	0	0.00%	0	0.00%
Missing	10		7		7	
Marital Status						
Married	610	100%	434	71.15%	467	76.56%
Widowed	57	100%	31	54.39%	33	57.89%
Divorced	1	100%	1	100.00%	1	100.00%
Separated	4	100%	3	75.00%	4	100.00%
Never Married	767	100%	546	71.19%	582	75.88%
Missing	1		0		0	
Birthplace						
Foreign	99	100%	63	63.64%	67	67.68%
US	1,243	100%	951	76.51%	1,020	82.06%
North Dakota	1,233	100%	886	71.86%	950	77.05%
Missing Birthplace	98		1		0	
Total	1,440	100%	1,015	70.49%	1,087	75.49%

Notes: Observations were first matched using name, sex, quarter of birth, and year of birth. The final column added birthplace as a matching variable. Rows titled 'missing' report the number of observations for which a variable was not reported.
Source: 1960 Census North Dakota Sample

6.2 Assigning PIKs to Children

Table 4 reports the PIK rates for children. When matching on first name, middle initial, sex, quarter of birth, and year of birth, the PIK rate for children is 72.2 percent. The PIK rate is highest for children born in the fourth quarter (78.0 percent) and lowest for children born in the third quarter (65.8 percent).

Table 4: Match Rates of Children by Demographic Characteristics

	Full Sample		Matched		Matched with Parent's Name		Matched with Parent's Names and Birthplace	
	N	%	N	%	N	%	N	%
Sex								
Male	369	100%	269	72.90%	297	80.49%	300	81.30%
Female	348	100%	249	71.55%	265	76.15%	267	76.72%
Missing	2		1		1		1	
Quarter of Birth								
Jan-March	176	100%	128	72.73%	136	77.27%	137	77.84%
April-June	182	100%	134	73.63%	145	79.67%	145	79.67%
July-Sept	202	100%	133	65.84%	149	73.76%	153	75.74%
Oct-Dec	159	100%	124	77.99%	133	83.65%	133	83.65%
Race								
White	712	100%	515	72.33%	559	78.51%	564	79.21%
Black	1	100%	0	0.00%	0	0.00%	0	0.00%
Missing	6		2		4		4	
Birthplace								
Foreign	8	100%	6	75.00%	6	75.00%	6	75.00%
US	711	100%	513	72.15%	557	78.34%	562	79.04%
North Dakota	695	100%	501	72.09%	544	78.27%	548	78.85%
Total	719	100%	519	72.18%	563	78.30%	568	79.00%

Notes: The sample is limited to children with a reported birthplace. First, they are matched using name, sex, quarter of birth, and year of birth. Then I add parents' names as a matching variable and birthplace. Each percent represents the match rate for a particular demographic category. Rows titled 'missing' report the number of observations for which a variable was not reported.
Source: 1960 Census North Dakota Sample

When using parents' first and last names, the PIK rate of children increases to 78.3 percent. The PIK rates increases by 7.6 percentage points for male children and 5.6 percentage points for children born in the fourth quarter. The PIK rate does not improve for foreign-born children when using parents' names.[12]

In the final columns of Table 4, I include birthplace information as matching variables. The overall PIK rate of children increases to 79.0 percent when using state of birth and the foreign-birthplace indicator. The greatest gain in the PIK rate occurs for children born in the third quarter, which increased by two percentage points from 73.8 percent to 75.8 percent. The PIK rate is highest for male children and children born in the fourth quarter, and the PIK rate remains lowest for children born in the third quarter and the foreign born.

7. 2010 Exercise

Comparing PIK rates does not tell us much about the quality of the matches. To investigate the match quality when you only have name, quarter of birth, and year of birth to match on, I reduced the identifiers on the 2010 census file to mimic the 1960 short form census. The altered 2010 PIK results were compared to PIKs assigned through CARRA's formal PVS process. Out of 302,103,352 individuals from the 2010 census processed, I am able to find a match for 192,143,629, or 63.6 percent. Of the observations assigned a match, I obtain a match identical to the formal PVS match for 178,343,078 individuals, or 92.8 percent.

A match rate of 63.5 percent is significantly lower than the match rate achieved for the 1960 sample. I limit my sample to individuals in the 2010 census who lived in the zip-3 area from which the North Dakota sample was taken to run a more comparable simulation. Using first name, middle initial, last name, sex, quarter of birth, and year of birth, I am able to match 80.2 percent to the Numident using the same parameters and passes used for the 1960 sample. Of those matched, 96.3 percent received the same match assigned by the formal PVS process. Therefore, it does appear that North Dakotans living in this region are more likely to receive a PIK relative to the entire 2010 census.

8. Conclusions

In this paper, I explore PIK rates of a small sample of the 1960 census. Although the sample is small, the results illustrate that it is feasible to match older census data to the Numident when name, quarter of birth, and year of birth are the only PII available for linking. Furthermore, the expected match rate is high, and evidence from matching the 2010 census to the Numident using the same procedure suggests the accuracy of the linkages is high as well. The results of this analysis demonstrate that match rates could be high for both the short form or the long form 1960 census.

[12] If native-born children were more likely to have their parents' names recorded in the Numident, we would expect this result. However, individuals who were 0-16 years old in 1960 are not any more or less likely to have their parents' names listed in the Numident.

9. References

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

Collins, W. J., & 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), 220-52.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association, 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, 34*, 141-56.

Goeken, R., Huynh, L., Lynch, T., & Vick, R. (2011). New Methods of Census Record Linking. *Historical Methods, 44*(1), 7-14.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the. *Journal of the American Statistical Association, 84*(406), 414-20.

Layne, M., Wagner, D., & Rothas, C. (2014). Estimating Record Linkage False Match Rate for the Person Identification Validation System (PVS). *CARRA Working Paper #2014-02*.

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

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

Mill, R. (2012). Assessing Individual-Level Record Linkage between Historical Datasets. *Preliminary Working Paper*.

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

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

Ruggles, S., Schroeder, M., Rivers, N., Alexander, J., & Gardner, T. K. (2011). Frozen Film and FOSDIC Forms: Restoring the 1960 U.S. Census of Population and Housing. *Historical Methods, 44*(4), 69-78.

Wagner, D., & 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. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-59.

Winkler, W. E. (1995). Matching and Record Linkage. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. A. Colledge, & P. S. Kott, *Business Survey Methods* (pp. 355-384). New York: J. Wiley.

Winkler, W. E. (2006). Overview of Record Linkage and Current Research Directions. *U.S. Bureau of the Census, Statistical Research Division Report, Research Report Series*.

10. Appendix

This appendix describes the sample used in the main analysis. Originally, 1,727 person records from the North Dakota 1960 census sample were delivered to CARRA for analysis. I limit the sample to the 1,400 observations that were linkable to the long form. The long form census contains birthplaces and I want to assess how using birthplace as an additional matching variable affects the match rate. To determine whether limiting the sample to observations with birthplace introduces bias, Table A-1 explores the presence and absence of birthplace data in the North Dakota 1960 census sample. Table A-1 shows that a lower percentage of individuals who are “other relatives” or “children” have birthplace reported, where as a larger percentage of those born in the third quarter of the year have a recorded birthplace.

Table A-1: Percent with and without a birthplace by demographic characteristic

	Birthplace		No Birthplace	
	N	%	N	%
Sex				
Male	732	82.99%	150	17.01%
Female	704	83.81%	136	16.19%
Relationship				
Head of HH	370	88.10%	50	11.90%
Wife of Head	293	86.43%	46	13.57%
Child	719	80.43%	175	19.57%
Other Relative	36	73.47%	13	26.53%
Non-Relative	20	86.96%	3	13.04%
Quarter of Birth				
Jan-March	384	82.23%	83	17.77%
April-June	346	83.78%	67	16.22%
July-Sept	390	85.71%	65	14.29%
Oct-Dec	320	81.63%	72	18.37%
Race				
White	1,429	83.37%	285	16.63%
Black	1	100.00%	0	0.00%
Marital Status				
Married	610	86.52%	95	13.48%
Widowed	57	89.06%	7	10.94%
Divorced	1	100.00%	0	0.00%
Separated	4	100.00%	0	0.00%
Never Married	767	80.57%	185	19.43%
Total	1,440	83.38%	287	16.62%

Notes: This table compares the percentage of observations with and without a recorded birthplace by demographic characteristic.
Source: 1960 Census North Dakota Sample

Table A-2 reports descriptive statistics for the entire raw sample, the sample with birthplace, and the group without birthplace. If observations are randomly assigned into the group with birthplace and the group without, we would expect the typical observable characteristics of those in the birthplace group to be similar to the non-birthplace group. A pooled t-test illustrates that this is not the case. The group with birthplace is older and has a higher percentage of heads of households and married individuals. The percentage of children and never married is lower for the birthplace group. These differences demonstrate that those who fall into the birthplace sample are not necessarily representative of the entire sample of 1,727 individuals taken from the 1960 short form.

Table A-2: Descriptive Statistics of those with and without Birthplace

	Full Sample		Birthplace		No Birthplace		Difference	
	Mean	N	Mean	N	Mean	N	NB-B	p-value
Age	27.988	1,726	28.890	1,439	23.467	287	-5.423	<.0001
Sex								
Male	0.511	1,727	0.508	1,440	0.523	287	0.015	0.658
Relationship								
Head of HH	0.243	1,727	0.257	1,440	0.174	287	-0.083	0.003
Wife of Head	0.196	1,727	0.204	1,440	0.160	287	-0.044	0.093
Child	0.518	1,727	0.499	1,440	0.610	287	0.111	0.001
Other Relative	0.028	1,727	0.025	1,440	0.045	287	0.020	0.059
Non-Relative	0.013	1,727	0.014	1,440	0.011	287	-0.003	0.643
Quarter of Birth								
Jan-March	0.270	1,727	0.267	1,440	0.289	287	0.022	0.433
April-June	0.239	1,727	0.240	1,440	0.233	287	-0.007	0.805
July-Sept	0.264	1,727	0.271	1,440	0.227	287	-0.044	0.120
Oct-Dec	0.227	1,727	0.222	1,440	0.251	287	0.029	0.290
Race								
White	0.993	1,727	0.992	1,440	0.993	287	0.001	0.905
Marital Status								
Married	0.408	1,727	0.424	1,440	0.331	287	-0.093	0.004
Widowed	0.037	1,727	0.040	1,440	0.024	287	-0.016	0.214
Divorced	0.001	1,727	0.001	1,440	0.000	287	-0.001	0.655
Separated	0.002	1,727	0.003	1,440	0.000	287	-0.003	0.372
Never Married	0.551	1,727	0.533	1,440	0.645	287	0.112	0.001

Notes: Age is estimated by 1960-birthyear.
Source: 1960 Census North Dakota Sample