HHS Public Access

Author manuscript

Annu Rev Sociol. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

Annu Rev Sociol. 2018 July ; 44: 19–37. doi:10.1146/annurev-soc-073117-041447.

Historical Census Record Linkage

Steven Ruggles, **Catherine Fitch**, and **Evan Roberts**

University of Minnesota

Abstract

For the past 80 years, social scientists have been linking historical censuses across time to study economic and geographic mobility. In recent decades, the quantity of historical census record linkage has exploded, owing largely to the advent of new machine-readable data created by genealogical organizations. Investigators are examining economic and geographic mobility across multiple generations, but also engaging many new topics. Several analysts are exploring the effects of early-life socioeconomic conditions, environmental exposures, or natural disasters on family, health and economic outcomes in later life. Other studies exploit natural experiments to gauge the impact of policy interventions such as social welfare programs and educational reforms. The new data sources have led to a proliferation of record linkage methodology, and some widespread approaches inadvertently introduce errors that can lead to false inferences. A new generation of large-scale shared data infrastructure now in preparation will ameliorate weaknesses of current linkage methods.

Introduction

Historical census data are indispensable for studying long-run change, since they provide the only record of the lives of millions of people over the past two centuries. In much of Europe and North America, investigators can access individual-level data based on original census manuscripts (Szołtysek & Gruber 2016). These documents record the name and characteristics of each individual counted by the census, organized into households. Used in isolation, however, these data have some significant limitations. Census records are cross-sectional snapshots, making it impossible to observe change over the life course or across generations. Furthermore, Census listings often lack sufficient information to investigate key research questions. Social scientists since the 1930s have turned to linking historical census records—both linking successive censuses over time and linking the census enumerations to other sources—to overcome these problems. The resulting research has profoundly influenced our understanding of the past.

Early record-linkage studies focused on social and geographic mobility in particular communities. Technological developments and new data sources since the late twentieth century have allowed for increasingly ambitious investigations. The new research spans diverse topics, including family transitions, the impact of early-life conditions on later outcomes, the consequences of slavery and reconstruction, the impact of policy interventions such as social welfare programs and educational reforms, and the complex interactions

between migration and labor markets. More broadly, the new research provides fundamental insights into the ways that industrialization, urbanization, immigration, and demographic transition have transformed society.

Census record linkage is complicated by the lack of a unique identifier across datasets; matching relies heavily on names, in combination with age, birthplace, and other characteristics. Census enumerators collected information from households and transcribed answers on the enumeration form; they usually spoke with only one person, the household head or spouse. Even this limited set of characteristics—names, ages and birthplaces—is likely to be slightly inconsistent across enumerations for many people and more likely to be imprecise or wrong for non-family members such as lodgers, and farm and household laborers. Further difficulties arise because of respondents with common names, women changing surnames upon marriage, immigrants anglicizing their names, and errors in transcribed data. This situation creates the potential for both false matches and missed matches. The algorithms scholars use to make linkages can inadvertently introduce errors that can lead to false inferences.

This essay describes the history of census record linkage and summarizes some of the fruits of this research. We begin with the history of community-level census record linkage from the 1930s to the 1980s. We then describe how the methods of historical census record linkage were successively transformed by new methods and sources. The introduction of indexes to the census in the late 1970s enabled construction of the first national linked samples. In the early 2000s, the advent of complete-count census databases stimulated a profusion of new methods for automated record linkage. We discuss some strengths and weaknesses of different approaches to automatic record linkage, and explain the potential for a new generation of shared data infrastructure to improve the reliability and replicability of research based on linked historical censuses. While research with linked census data is international in scope, most of the published and forthcoming literature we discuss is from the United States. Beyond the work discussed in this review the growing availability of linked datasets in Canada, Great Britain, Ireland, and Scandinavia should spur even more research from these countries in forthcoming years.

Census Record Linkage at the Community Level, 1935–1995

Malin (1935) used state and federal censuses to link Kansas farm operators across 13 censuses taken between 1860 and 1935, and found extraordinary levels of turnover. Challenging the “safety valve” thesis of Frederick Jackson Turner (1893), Malin found that migration to frontier communities was lowest when the economy was poor. Twenty-five years later, census record linkage was the central underpinning of the “New Social History,” which emphasized a study of ordinary people, everyday life, and “history from below”. New social historians emphasized that the census was the only surviving record of many peoples’ lives. Linking the population of Trempealeau County, Wisconsin across four censuses (1850–1880), Curti (1959) also documented extraordinary levels of population turnover, with some three-fourths of inhabitants disappearing each decade. Thernstrom (1964) linked men across censuses from 1850 through 1880 together with tax records and local savings bank records to assess social mobility in Newburyport, Massachusetts. Thernstrom

challenged the "rags-to-riches myth," arguing that only a small percentage of unskilled laborers ever managed to move up the economic hierarchy.

In the late 1960s and 1970s, dozens of social historians followed the path laid out by Malin, Curti, and Thernstrom. Most of these focused on cities or towns rather than rural areas, such as Atlanta (Hopkins 1968), Boston (Knights 1971), South Bend, IN (Esslinger 1975), Hamilton, Ontario (Katz 1975), and Kingston, NY (Blumin 1976). The most prominent investigation was the Philadelphia Social History Project (Hershberg 1976), which digitized 2.5 million census records and linked them to other sources in the Philadelphia area such as church and school records.

These studies shared the same basic methodology. Starting with a base census year and town or county, the investigator identified a population of men to follow, because men's economic status determined the family's economic status and men's names did not change upon marriage. Most investigators then manually searched the microfilmed census listings for the same locality in the subsequent census year to identify the same individuals. In some cases, investigators used sorting machines or computers to alphabetize the names. Men successfully located in both censuses were designated "persisters," a term coined by Malin. In both rural and urban settings, the percentage of persisters was typically under 40% (e.g., Blumin 1976; Bogue 1963; Faragher 1986; Knights 1971; Thernstrom 1964)

What happened to the non-persisters? Many died between censuses; others were missed because of errors[1] ; and the rest migrated out of the community. Lack of information about the fate of outmigrants limits the potential of the community studies for generalizations about life-course change. Thernstrom's pessimistic reading of the prospects for economic mobility in nineteenth-century America was based on his speculation that the movers represented a "floating proletariat" who had little prospect of success (Thernstrom 1964, 1973). As Ferrie (1999) and others have pointed out, however, there was no real basis for that speculation; subsequent research showed that men who left Newburyport actually fared better than those of similar status who remained behind (Herscovici 1998).

Excluding most of the population from the analysis not only damaged studies of economic mobility, it also compromised virtually all generalizations based on the linked community datasets. We now know that the persisters in the community studies were highly selected and had markedly different characteristics from the unlinked. Accordingly, conclusions about life-course changes based only on tracing persisters should be qualified as applying to an unrepresentative minority of the population.

National Linked Samples Based on Name Indexes, 1975–2000

A new approach became possible with the release of Soundex indexes to the U.S. census. These name indexes were prepared under the Works Project Administration between 1935 and 1941, and were intended to allow individuals applying for old age benefits to prove their

[1]In addition to misreported data--errors from the enumerator or the household responder—there are also illegible entries due to poor handwriting, damage to the manuscripts, or poor microfilming. The enumerator also missed some individuals or entire households; estimates for nineteenth and twentieth century net underenumeration range from 3.8% to 6.6% (Hacker 2013).

age without a birth certificate using information reported in earlier censuses (Prechtel-Kliskens 2002). Soundex is a phonetic encoding system used to avoid the problem of variant name spellings. In 1973, the National Archives opened the 1900 census listings with a complete Soundex index to persons engaged in "bona fide research," and the scholarly community began to take notice (Jensen 1974; Stephenson 1974, 1975; Thernstrom 1975).[2]

The Soundex indexes were the underpinning of the National Panel Study (Guest 1987), the first national record linkage project. The National Panel Study drew a sample of 10,000 males aged 5–14 and 25–34 from the 1880 census, and used the 1900 Soundex to trace them to the 1900 census. A review committee evaluated potential links and made judgements of correct or near matches on age, birthplace, and presence of the same family members in the two censuses. Cases with multiple matches were eliminated. The Soundex is organized state-by-state and was available only on microfilm so lookups were expensive. Linking interstate migrants would have required searching every state, and that was not financially feasible. The investigators did search the most frequent destinations for out-migrants from each state, but the effort was "laborious and disappointing" (Guest 1987: 65); despite the availability of Soundex, the investigators located only 160 interstate migrants.

In 1982, a genealogical publisher completed a name index of the 1850 census (Jackson 1982). Steckel (1988) exploited this tool to construct a national panel for the 1850–1860 period consisting of free male household heads with children aged 10 or older in 1860. Like the Soundex, the 1850 index was organized by states, so tracing interstate migrants was still costly. Steckel used the birth states of children in 1860 to decide which states to search in 1850, eliminated names that appeared more than 10 times, and confirmed matching names using age, birthplace, and the presence of the same family members.

A genealogical index of the 1860 census appeared in 1992, and for the first time, the data became available as a national machine-readable file, instead of as separate indexes for each state (Jackson 1992).[3] In 1993, investigators at the University of Minnesota released IPUMS, an integrated series of public-use microdata including samples of historical U.S. censuses (Ruggles & Menard 1995). Ferrie (1996) capitalized on both of these new resources to link men in the IPUMS sample of the 1850 census forward to the 1860 index. Unlike previous investigators, Ferrie used a set of formal rules to guide his record linkage, and he was able search nationally for interstate migrants. Using phonetic encoding, he identified all possible matches in the 1860 index. He dropped cases with 10 or more potential matches, and manually looked up the remaining potential matches on microfilm of the census listings. He then eliminated potential matches with a substantial age discrepancy, mismatching birth state, or mismatching family members. If two or more matches remained, he chose the one that was the closest match on age. Ferrie's rule-based linkage algorithm had a profound impact on subsequent research. In most respects, Ferrie's rules approximated the guidelines established by Guest and Steckel, but they eliminated the role of judgement.

[2]In the United States census enumerations are released to the public 72 years after they are taken. A partial Soundex index of the 1880 census had apparently been available at the main branch of the National Archives since the mid-1960s, but no references to it appear in the scholarly or genealogical literature until 1974.

[3]The national file was constructed by Joseph Ferrie, who interleaved machine-readable state files to create a national file (J.P. Ferrie, personal communication).

The Advent of Automatic Record Linkage

Between 1982 and 1999, the Church of Jesus Christ of Latter-day Saints organized more than 20,000 volunteers to digitize over 80 million records, comprising the entire 1880 census of the U.S. and the 1881 censuses of Canada and Great Britain. As the projects were nearing completion, the Church negotiated agreements with academic organizations in each country to convert the transcriptions into cleaned and coded databases suitable for scientific research.[4] At the same time, investigators in Norway and Iceland were constructing complete machine-readable nineteenth-century census enumerations. To ensure compatibility of the databases, the investigators from the six countries formed a collaboration called the North Atlantic Population Project (NAPP), and negotiated common data formats and classifications. The initial datasets for each country were released at nappdata.org in 2003 and 2004 (Roberts et al. 2003).

The availability of complete census enumerations in machine-readable form made it feasible to conduct automatic census record linkage. Within a few months of the first NAPP data release, Ferrie (2004) constructed a set of five linked samples using the 1880 complete count data for the U.S. linked to the IPUMS 1% samples for 1850, 1860, 1870, 1900, and 1910. The complete-count data allowed the abandonment of the genealogical indexes, which meant that cases no longer had to be manually verified by locating them on microfilm. Ferrie's new linking algorithm dropped phonetic encoding of names and instead linked only those cases with an exact match or very close spelling variation on name.[5] The algorithm also required an exact match on birthplace (and parental birthplace if available) and a birth year that differed by no more than three years. Ferrie dropped cases with more than one match, and used weights to correct the sample for biases with respect to region of residence, urban residence, migration, occupation, and other variables. The new Ferrie method was fully automated and clearly specified, which makes it replicable, usable by other researchers, and practical to evaluate for errors and bias. Following Ferrie's approach, Long (2005) linked 28,000 men from a sample of the 1851 census of England and Wales to the complete-count British census of 1881.

The linked samples based on the first complete-count datasets yielded striking and controversial findings. Ferrie (2005) found that nineteenth-century U.S. migration was extraordinarily high and was positively associated with upward economic mobility. This finding is consistent with Turner (1893), but it undercuts the arguments of generations of sociological theorists that ever-increasing mobility in the twentieth century contributed to reduced family cohesion, social dislocation, disrupted schooling, and health impairment (e.g., Litwak 1960, Parsons & Bales 1955). Ferrie's findings also contradicted the interpretation of Thernstrom (1964) and the other new social historians that migrants faced poor economic prospects in the nineteenth century.

[4]These were IPUMS at the University of Minnesota, the U.K. Data Archive at the University of Essex, and the Department of Demography at the University of Montreal.

[5]A close spelling variation was defined as a score of 30 or less using the SPEDIS algorithm in SAS. Ferrie also standardized diminutives of first names (e.g., Joe and Bill) and eliminated cases in which one census listed first name with only an initial.

Long and Ferrie (2007, 2013) built on these results by comparing intergenerational occupational mobility in Britain and the U.S. between the mid-nineteenth and late-twentieth centuries. According to Long and Ferrie, economic mobility was far higher in the U.S. than in Britain in the nineteenth century, but U.S. economic mobility declined and converged with British levels by the late twentieth century. These findings challenged 50 years of sociological research arguing that relative social mobility in industrialized countries has been constant or increased over time. Critics raised concerns about the reliability and representativeness of the linked data, as well as the statistical measures used to assess occupational mobility (Hout & Guest 2013, Xie & Killewald 2013).

IPUMS Linked Representative Samples

The IPUMS group adopted a different strategy for exploiting the new complete count 1880 data for automatic record linkage (Ruggles 2002). Most prior record linkage efforts, they argued, had focused too much attention on the percent of persons missed by record linkage (Type II errors), and not enough about the percent of false links (Type I errors). Failing to identify links can lead to selection bias, but investigators can measure that bias and largely correct for it through weighting. False matches, by contrast, can introduce systematic bias into many kinds of analysis. Suppose, for example, that an investigator seeks to measure migration. Falsely matched cases ordinarily appear as migrants, since two falsely-linked individuals seldom reside in precisely the same place. False matches lead to systematic upward bias in migration rates, occupational mobility, and all kinds of family transitions. To address this problem, record linkage should not attempt to maximize the number of links, but rather should maximize the accuracy and representativeness of the links. The IPUMS Linked Representative Samples (IPUMS-LRS) aimed to minimize selection bias with respect to key transitions such as migration or widowhood. This required using a limited set of characteristics not expected to change over the life course: name, birth year, sex, and birthplace.[6] With the exception of Ferrie (2004), previous record linkage studies had used other information on the record—such as place of residence, occupation, or names of spouses—to resolve ambiguities.

IPUMS-LRS linked datasets on a probabilistic basis using the Jaro-Winkler string comparison metric developed by the Census Bureau (Porter & Winkler 1997) and a machine-learning tool known as a Support Vector Machine (Chang & Lin 2011, Christen 2008). The probabilistic machine-learning software compared every person of a given birthplace and sex with every other person that shared those characteristics and predicted the probability of a true match based on similarity scores of several features, such as spelling of first name and last name, first and middle initials, phonetic name codes, name commonness, and age. The machine-learning software was “trained” with a set of hand-linked data developed by IPUMS staff.

By adopting conservative linking thresholds, IPUMS-LRS minimizes false matches (Goeken et al. 2011, Ruggles 2011). Under the IPUMS-LRS linking procedure, whenever more than a single potential match was found, all potential matches were eliminated from consideration.

[6]The datasets excluded women who married during an interval between censuses because they ordinarily changed their surnames.

To weed out false matches, IPUMS-LRS developed two models. The "loose" model was designed to maximize the number of potential links. The "tight" model was more selective, and established matches only where the fit was extremely close. Links were designated as true only if there was one and only one positive link in both models (Goeken et al. 2011: 10). This approach sacrifices valid links to minimize false links and maximize representativeness. The loose model excludes cases with an observed possibility of choosing the wrong match, and the tight model excludes cases with any significant discrepancies in name or age. The samples were weighted to approximate the characteristics of the potentially-linkable with respect to family relationship, birthplace, age, size of place, and occupation. Evaluation using consistency checks suggests that IPUMS-LRS has a very low rate of false matches (Goeken et al. 2011).

The IPUMS-LRS was first released in 2008. Unlike previous record linkage projects, IPUMS-LRS was not designed to investigate a particular research question; rather, the project was conceived from the outset to be general-purpose shared data infrastructure. IPUMS-LRS constructed 28 linked public use files, including separate files for men, women, and married couples linked from the U.S. 1880 complete-count data to the IPUMS samples for 1850, 1860, 1970, 1900, 1910, 1920 and 1930 (Goeken et al. 2011, Ruggles 2011). Several investigators adopted variants of the IPUMS-LRS approach to create linked samples for the other countries with complete digital census enumerations: England and Wales 1851–1881, Norway 1865–1900, Canada 1871–1881, and Sweden 1880–1890 (Ruggles et al. 2011, Antonie et al. 2014, Wisselgren et al. 2014).

IPUMS-LRS yields lower estimates of migration and intergenerational occupational mobility than do the Long-Ferrie linked samples (Baskerville et al. 2014). This result suggests that the extremely high geographic and occupational mobility Long and Ferrie found in the nineteenth-century U.S. may be partly ascribed to false matches. The only cases Long and Ferrie discard are those with multiple perfect or near-perfect matches. In some cases, however, the perfect match is false, and the true match has a slightly different spelling. The loose IPUMS-LRS model broadly discards cases with any competitors that have a possibility of being true matches.

Many studies have used IPUMS-LRS to go beyond estimation of levels of geographic and economic mobility. For example, Saperstein and Gullickson (2013) explored the fluidity of black and mulatto racial categories between the 1870 and 1880 censuses. Rauscher (2016) used variation in the timing of compulsory school attendance laws to study the impact of those laws on differentials in occupational mobility. Paradoxically, the immediate impact of compulsory attendance was to reduce occupational mobility, possibly because schools in working-class districts were unprepared for the sudden influx of new students. Bloome and Muller (2015) used IPUMS-LRS to argue that high levels of tenant farming were associated with high levels of both marriage and divorce among African-American in the postbellum South. Bleakley and Lin (2012) demonstrated that workers change occupations and industries less often in densely populated areas, and that this relationship has remained stable since the nineteenth century.

Historical Census Record Linkage in the Era of Big Data

The past few years have witnessed astonishing advances in the development of complete-count historical census data, and these new resources have created the potential for far more powerful linked historical censuses. After 2000, Ancestry.com, FamilySearch, and FindMyPast.com all began heavily investing in digital transcriptions of historical censuses. Social scientists began using these resources as soon as they became available, using manual searching on the genealogical websites to locate populations of interest. Some investigators started "scraping" data from the websites using scripts to make repeated queries. The genealogical organizations disapprove strongly of this practice; for some of the organizations, scraping is a direct violation of the terms of service, and they all monitor closely and protect proprietary control of their own data. Eventually, academic organizations in the U.K. and the U.S. secured agreements to disseminate legally the genealogical indexes to academic researchers. Under those agreements, anonymized datasets would be made freely available to researchers, and restricted datasets including names would be made available under contracts that safeguard the proprietary interests of the genealogical organizations.

Researchers now have access to nine complete enumerations of the United States (1850–1940) and to seven for Britain (1851–1911) (Ruggles 2014, Schurer & Higgs 2014). In the same period, national archives and social science research organizations in Canada, Denmark, Iceland, Ireland, Norway, and Sweden launched efforts to digitize censuses. These investigators have produced complete microdata for 17 national censuses taken between 1703 and 1910, and more are on the way (Antonie et al. 2015, Ruggles et al. 2011, Thorvaldsen 2007).[7]

The IPUMS project released the first version of the complete-count U.S. data for years other than 1880 in late 2013. Investigators immediately began linking them to one another and to other sources, finding creative ways to address previously intractable questions. At this writing, nearly 500 researchers are working on 188 research projects with the restricted U.S. complete-count data with names, and 148 of those projects involve record linkage. This count represents just a fraction of current research linking historical census records; it excludes, for example, all the projects using exclusively British, Canadian, Nordic, or Irish data. There is a profusion of research topics represented among these linkage projects using complete-count data. The paragraphs that follow describe a small sampling of this research, some of it still in the early stages; these examples convey the range of inventive research strategies investigators have developed in just the past few years.

New studies of social and economic mobility

Feigenbaum (2017) matched men in the 1915 state census of Iowa—the first U.S. census to include a question on income—to the 1940 census to assess intergenerational elasticity of income, and measured higher intergenerational mobility than is found in recent data.

[7]The anonymized versions of all these datasets data are available through ipums.org or nappdata.org. For access to names, researchers must apply to the U.K. Data Archive at the University of Essex for the British data, and to IPUMS at the University of Minnesota for the U.S. data. The full Nordic databases, including names, are available at nappdata.org, and the Irish data will be available soon.

Modalisi (2017) carried out an ambitious analysis of long-run changes in intergenerational occupational mobility by linking seven Norwegian censuses spanning 1865 to 2011. In sharp contrast to the finding of Long and Ferrie (2013) that occupational mobility declined in the U.S., Modalisi found a steady and substantial trend towards increasing mobility in Norway. Eli, Salisbury and Shertzer (2016) linked Kentucky civil war pension records to the censuses of 1860 and 1880. They used these data to characterize the impact of the war on occupational mobility of Union and Confederate veterans, finding that confederate veterans who remained in Kentucky had better economic outcomes than did union veterans, whereas the opposite was true among those who migrated out of state.

These new data provide opportunities to go beyond two-generation studies of male economic opportunity. Ward (2017) examines occupational mobility over three generations, by linking foreign-born grandfathers in the 1880 census to sons in 1910 and grandsons in 1940, finding that the skill level of the grandfather had a powerful impact on the skill level of his grandchildren. Ferrie, Massey and Rothbaum (2016) extend multigenerational analysis even farther, by linking the 1910, 1920, and 1940 censuses to internal versions of Census Bureau surveys from the 1970s to the 2010s. By using Social Security records, the authors were able to obtain maiden names, allowing intergenerational record linkage for women as well as men. The study found significant grandparent effects on educational attainment, but no significant effects of great-grandparents. The Swedish linked censuses also cover women as well as men, allowing Dribe, Eriksson, and Scalone (2017) to uncover a strong positive association between economic mobility and migration among women between 1880 and 1900. Salisbury (2017) linked Civil War pension records to Union Army widows, and found that those who received a pension had substantially lower odds of remarriage, probably because the pensions made non-marriage a viable option.

Immigration and internal migration

Abramitzky, Boustan and Eriksson (2012) link the Norwegian census of 1865 to the Norwegian and U.S. censuses of 1900. They compared the occupational outcomes of brothers when one migrated to the U.S. and the other remained in Norway, and estimate an economic return to migration of 70 percent, which is comparatively low by recent standards. Connor (2017) linked the Irish census of 1901 to the U.S. censuses of 1920 and 1940 to evaluate how local conditions in Ireland affected the later life outcomes of Irish immigrants and their children. Linking the 1910 census to the 1930 census, Collins and Wanamaker (2017) examined the selection and sorting of Southern black and white migrants in the Great migration and found distinctive race differences in migration that cannot be ascribed to initial characteristics. In particular, black migration was more responsive to labor demand, and blacks were more likely than whites to head for manufacturing centers.

Impact of early-life health conditions and environmental exposures on later outcomes

Beach et al. (2016) examined the impact of typhoid exposure in childhood (1889–1899) on later-life income and education, finding that eliminating typhoid from a city increased average later-life earnings by one percent and educational attainment by one month. Warren et al. (2012) matched pairs of brothers, one of whom was identified as sick and the other as healthy in the 1880 census, to the 1900, 1910, 1920, and 1930 censuses. In a similar study,

Karbownik and Wray (2016a) matched London hospital admissions for children between 1874 and 1901 to the British censuses of 1881, 1891, 1901, and 1911. All these studies uncovered profound adverse consequences of childhood sickness for later-life occupational attainment.

Many studies leverage environmental exposures to conduct natural experiments. Karbownik and Wray (2016b) matched World War I Draft Registration Cards with place of birth information to late-nineteenth century hurricane paths and the 1940 U.S. census to assess the impact of fetal and early childhood exposure to stress caused by hurricanes. Parman (2015) matched World War II military enlistment records to the 1930 census to show how exposure in utero to the 1918 influenza pandemic not only had profound consequences for later life health and educational attainment, but also positively affected the educational attainment of siblings, presumably because of a reallocation of household resources. Ferrie, Rolf and Troesken (2012) estimated lead exposure in water supplies for children in the 1930 census, and linked them with intelligence test scores for World War II enlistees; they found a powerful inverse association between lead exposure and test scores.

Kin relationships beyond the household

Historical censuses and census-like listings have been fundamental to understanding global household and family change, because they allow consistent comparison of living arrangements across many countries and hundreds of years (Laslett and Wall 1972; Ruggles 1994, 2008). For the past half-century, social theorists have frequently objected that because such sources cannot identify kin relationships beyond the household, they often miss important relationships (e.g. Sussman 1959; Litwak 1960). Record linkage provides a means to overcome this limitation. Jennings, Sullivan, and Hacker (2012) used linked genealogical information from the Utah Population Database to identify grandparents and other relatives to assess the intergenerational transmission of reproductive behavior. Hacker and Roberts (2017) identified neighboring parents in the complete-count 1880 United States census to measure the impact of kin propinquity on fertility. With the development of new data infrastructure, these kinds of analysis will become easier. In the next few years the IPUMS-MLP project (discussed in greater detail below) will develop a multi-generational panel dataset from United States historical census records. As family members are traced across generations, they will be linked to common grandparents, thus allowing identification of aunts and uncles, cousins and siblings, regardless of where they live.

Reliability of Automated Census Record Linkage

The proliferation of census linking methodology is nearly as great as the proliferation of research topics. We are witnessing the wild west of record linkage: almost every new study introduces some new variant in methodology. A disadvantage of this methodological diversity is that one often cannot compare results across studies, and it can be difficult to replicate studies. Of even greater concern, however, is the great variation in the reliability of the methods. As noted earlier, linked historical censuses are subject to two sources of error, either of which can lead to incorrect inferences. Errors of omission (Type II errors) occur when records are not matched due to errors in the data or because multiple possible matches

exist and it is impossible to determine which one is correct. Errors of commission (Type I errors) occur when records are matched incorrectly. Either type of error can produce biased estimates.

Errors of omission

Missed matches can result in incorrect inferences by introducing selection bias. For example, if only the wealthy are matched, linked datasets will yield inaccurate estimates of any behaviors associated with wealth. In most cases, it is comparatively simple to measure the rate of missed matches. The first step is to estimate the population at risk to be linked, here termed the potentially-linkable population. The potentially-linkable population is defined as the population alive and present in both sources being linked. When linking two censuses from different years, the potentially-linkable population can be approximated by counting persons alive and present at the more recent census who were old enough to be present at the initial census year and who are either native-born or who immigrated before the initial census year. By measuring the potentially linkable population at the second observation, we eliminate cases that are unlinkable because of death or emigration. For some census linkages, the potentially-linkable population can be measured directly from the census; in other years, it must be estimated using data from other sources (Goeken et al. 2011).

The potentially-linkable population is the denominator needed tor measuring linkage rates, and the numerator is the number of successful matches. Most linkage studies calculate linkage rates as the percentage of persons in the initial census year who are linked to a subsequent census. This approach necessarily understates linkage rates and overstates selection bias because it ignores death and emigration.

As long as the more recent census includes information on age, birthplace, and year of arrival for immigrants, we can identify the potentially-linkable population at the individual level.[8] This enables straightforward tests of representativeness: investigators can compare the linked cases with potentially linkable cases, and directly observe the levels and significance of discrepancies. Analysts can then use these data to correct for non-representativeness through weights (e.g., Goeken et al. 2011, Mill & Stein 2016). Such weights cannot correct for unobserved characteristics, but to the extent that unobserved characteristics are associated with observed characteristics, it will mitigate the problem.

Errors of commission

Some methods have high rates of false matches, and this has the potential to result in misleading inferences. As noted, false matches exert systematic upward bias when measuring changes across linked sources, such as changes in place of residence, occupation, or marital status. Transition rates are systematically biased by Type I errors because the false matches are linking two different people, and different people usually have different characteristics.

[8]All twentieth-century U.S. censuses include information on year of arrival except for 1940–1960, the period when immigration was lowest.

Measuring the rate of false matches in a linked dataset is difficult. Some analysts have evaluated false positive rates of different methods by implementing those methods for datasets where the true links are known. For example, Massey (2017) evaluated several record linkage protocols by using them to link respondents to the Current Population Survey with Social Security records. Those records had already been linked by Social Security number, so the true links were known with high confidence. Similarly, Bailey et al. (2017) used genealogically-linked data as a "truth deck" to evaluate several record linkage strategies; although the genealogical datasets may include some errors, they are presumed to be comparatively high quality. Machine-learning linking algorithms provide another approach. Machine learning generally requires "true" matched data to train the software. That training data can be divided into parts, with one part used for training the linking software and another part used to evaluate accuracy of the trained software (Feigenbaum 2016, Parman 2015b).

Because historical censuses lack unique identifiers such as Social Security numbers, names play a critical role in record linkage. Names are frequently misspelled in census records and other sources, so virtually all linking methods incorporate some mechanism for matching names that have slight variations. There are two general approaches: phonetic encoding and string comparison. The main phonetic encoding systems used for census record linking are Soundex (the basis for the WPA census indexing discussed above), the New York State Identification and Intelligence System (NYSIIS), Metaphone, Double-Metaphone, and Phonex. NYSIIS is a refinement of Soundex, with better accommodation of European and Hispanic pronunciations. The Metaphone algorithm includes special rules for handling spelling inconsistencies and combinations of consonants, and Double Metaphone adds an alternate phonetic code to reflect non-English pronunciations. Phonex combines elements of Soundex and Metaphone to maximize true matches (Lait & Randell 1996).

Phonetic encoding is easy and inexpensive, but it is also a blunt instrument. Phonetic codes were designed not for final matching, but to allow manual lookups that would be manually confirmed by examining the fully spelled-out names. The phonetic coding schemes group together names that any human would regard as distinctly different. The Mill and Stein (2016) implementation of NYSIIS, for example, categorizes the following pairs of names as perfect matches: John and James; Hart and Heyward; and Sales and Schools. Used by itself, therefore, phonetic encoding has an intrinsically high potential for false matches. Evaluations of the accuracy of record-linkage methods provide strong evidence that relying entirely on phonetic encoding can have extremely high false match rates, in the range of 20% to 70% of links (Bailey et al. 2017, Massey 2017).

The alternative to phonetic coding is string comparison algorithms, which measure the number of differences between two character strings, yielding a statistic describing the degree of similarity of any two names. There are several measures in widespread use; they vary by such factors as the relative weight they place on transpositions, substitutions, string length, and position within the string of the discrepancies. The leading string comparator is Jaro-Winkler, which was developed at the Census Bureau and is optimized for comparisons of names (Christen 2012, Porter & Winkler 1997). A few linkage algorithms use the Levenshtein Distance, which is simply the minimum number of edits needed to transform

one string into another; the Damerau-Levenshtein distance, which also incorporates swapping of adjacent letters; or the SPEDIS algorithm which is part of the SAS statistical package (Hall & Dowling 1980, Roesch 2012).

String comparators are expensive to use with large datasets, since it is necessary to compare every name in one dataset with every other name in the other dataset to calculate which ones are the closest matches. When there are millions of names to compare, this demands very large computing capacity. As a solution, many analysts use phonetic encoding in combination with string comparison algorithms. By first subdividing names according to phonetic codes—a procedure known as blocking—the number of string comparisons needed can be dramatically reduced without losing a significant number of "true" matches (Mill 2013).

Because many names occur frequently, there are often multiple potential matches from which to choose. When multiple matches occur, no more than one can be correct. Many investigators try to choose the "best" match among multiple matches, but such techniques often allow false matches to slip through. The least reliable methods are those that seek to make use of all possible matches. For example, Nix and Qian (2015) randomly chose one match whenever their algorithm produced multiple tying matches. This gave them a very high match rate but an extremely high false-match rate, estimated by Bailey et al. (2017) at between 52% and 70%. Nix and Qian used their linked data to argue that 19% of black men "passed" for white at some time during their lives, and that 10% then reverted to identifying as black. In view of the high potential for false matches in the Nix and Qian methodology, this result should be viewed with skepticism.

Other methods suffer because they define multiple matches too narrowly. For example, to link the 1865 Norwegian census with the 1900 U.S. census, Abramitzky, Boustan and Eriksson (2012, 2013, 2014) assigned matches in three stages. First they linked all unique perfect matches on both the NYSIIS coded name and birth year, and in subsequent stages they allowed one- and two-year age discrepancies. Unfortunately, many true matches do not match precisely on year of birth. Censuses rarely ask directly about birth year; it usually must be inferred by subtracting year of enumeration from age at last birthday. This calculation is highly sensitive to the date of the enumeration, which varies from census to census. Moreover, ages were often approximated, especially for non-family such as boarders or farm hands. If the true match is off by just one year, and there is a false match with a perfect NYSIIS name code and birth year, Abramitzky et al. will assign the incorrect match. Using several different datasets, Bailey et al. (2017) estimate a false match rates between 25% and 38% from the Abramitzky et al. method.

Abramitzky, Boustan and Eriksson (2014: n. 27) recognized the potential for false matches to bias their results, and to evaluate the potential impact of the problem they constructed a "restricted" sample that discarded any links that were not unique within a five-year age band. The study found that the wage premium for long-term immigrants was twice as large in the restricted sample as in the main sample. This result is consistent with a hypothesis that false matches in the main sample diluted the observed wage premium. The most

conservative approach for minimizing false matches is to discard all cases with multiple potential matches.[9]

There is a trade-off between false matches and missed matches: steps taken to reduce false matches will usually increase missed matches. Feigenbaum (2016) developed a machine-learning linking protocol based on IPUMS-LRS. Instead of a Support Vector Machine, Feigenbaum used a probit model, where the independent variables were similar to those used in IPUMS-LRS. A case was considered matched if its probit score both exceeded a minimum linking threshold, and exceeded the second-best match by a minimum gap threshold. As Feigenbaum recognized, the critical determinant of performance of the method is the level chosen for the two thresholds. Feigenbaum opted to tune the parameters to balance the false positive matches with the false negative matches. By balancing the level of false matches with the level of missed matches, the analysis obtained a higher number of links but also a much higher rate of false matches compared with IPUMS-LRS, estimated by Feigenbaum (2016) at almost 15% and Bailey et al. (2017) at between 14% and 37%.

Equal weighting of false matches and missed matches heightens the risk of invalid inferences. In most instances, false matches are considerably more problematic than are missed matches. False matches introduce systematic upward biases in transition rates, such as migration rates, economic mobility, family transitions, or fluidity in racial identification. It is extremely difficult to measure the level of false matches in a linked dataset, and it is difficult to detect and adjust for the biases they introduce. By contrast, the rate of missed matches is comparatively easy to measure, and it is comparatively easy to detect the biases they introduce and mitigate those biases through weighting. Depending on study design, false matches do not necessarily invalidate conclusions, but caution is essential.[10] Investigators using linked historical census data should carefully consider the potential impact of false matches on their analyses, and adopt methods shown to minimize the problem.

The Future of Linked Historical Censuses

This is an exciting moment for historical census record linkage. New studies using linked census data are appearing virtually every week, often addressing hitherto intractable problems in novel and creative ways. At the same time, the expansive range of roll-your-own linkage methods that investigators are using in these studies raises serious concerns. There are no standards for historical census record linkage. Few studies measure errors of omission

[9]Multiple matches—where one match is correct and one or more are false—do contain a signal of information, and it may be possible to extract that signal without introducing bias. Statisticians have proposed methods that may allow unbiased estimation using multiple matches in regression analysis (Chambers 2009, Lahiri & Larsen 2005, Scheuren & Winkler 1993, Hof and Zwinderman 2014, Goldstein, Harron, and Wade 2012). Developed with simulated data, these methods include assumptions that are highly unrealistic for historical census data. The limited empirical testing suggests that the methods yield only marginal improvements to estimates, even using very high-quality data (Chipperfield and Chambers 2015, Dalzell and Reiter 2017). Until we have clear empirical validation of the statistical methods for correction for false marches, the safest course is simply to keep false matches to a minimum.

[10]When using measures other than transition rates, false linkages introduce noise that may obscure relationships, but this sometimes just makes estimated effects more conservative. Consider, for example, the findings of Ferrie, Rolf, and Troesken (2012) that exposure to water-borne lead in childhood strongly affected scores on Army General Classification Tests. False matches would attenuate the strength of that association, but there is no plausible mechanism whereby false matches could produce the association. More broadly, except when measuring levels of transitions between censuses, false matches may blur results but will not necessarily invalidate conclusions.

(Type II errors) relative to the potentially linkable population, and even fewer studies weight their linked samples to match the characteristics of that population. Also disconcerting is that the rate of false matches may be intolerably high for several widely-used methods, especially those that rely on phonetic classifications for the final match. Because of the wide variety of methods currently in use, it is impossible to compare statistics across studies, and replication is more difficult than it would be if methods were more standardized.

Five major new census record linkage infrastructure projects are now underway will help address this problem. Like the IPUMS-LRS project described earlier, these projects are not being developed solely to address particular research questions. Rather, they are being designed as general-purpose shared data infrastructure that will be applicable to many research problems, and will allow investigators to use linked historical census data without having to cobble together a purpose-built dataset. The investigators on these five projects are meeting regularly to ensure that their efforts are compatible and complementary. The following paragraphs briefly summarize the goals of each project.

- **Linking 1940 U.S. Census Data to Five Modern Surveys.** John Robert Warren is leading a project to link records from the 1940 census to records for respondents to the Health and Retirement Study (HRS); the Panel Study of Income Dynamics (PSID); the Wisconsin Longitudinal Study (WLS); the National Social Life, Health, and Aging Project (NSHAP); and the National Health and Aging Trends Study (NHATS). These ongoing longitudinal studies are the cornerstones of America's data infrastructure for interdisciplinary research on aging and the life course, including topics such as physical and mental health, disability, and well-being; later-life work, economic circumstances, and retirement; and end-of-life issues. This project will add critical information about social, economic, family, neighborhood, and environmental circumstances in childhood and young adulthood and allow researchers to examine the long-term impacts of these early-life circumstances and to understand how later-life outcomes are the result of cumulative life-course processes.
- **Census Longitudinal Infrastructure Project (CLIP).** CLIP is a major infrastructure project established by the Census Bureau's Data Stewardship Executive Policy Committee in August 2014 and is housed in the Census Bureau's Center for Administrative Records Research and Applications (CARRA) (Alexander et al., Johnson et al. 2015, Massey 2014). The goal of the project is to develop a general framework for longitudinal analysis of administrative and statistical records. The 1940 census is the centerpiece of the CLIP framework. The 1940 census—the first to provide such key indicators as educational attainment and income—provides the baseline population for constructing millions of life histories by linking forward to administrative records. To date, CLIP has positively identified 72% of children age 0–9 who appear in the 1940 census. CLIP data are already allowing investigators to understand the origins of later life outcomes among people who were children in

1940; several such research projects based on CLIP are already underway in Federal Statistical Research Data Centers (FSRDC).

- **American Opportunity Study (AOS).** The AOS team is building data infrastructure within CARRA to examine long-term trends in equality and the effect of program participation. Their goal is to combine data from multiple sources in ways that allow for more comprehensive answers than are possible with survey and experimental data alone (Grusky et al. 2015). Hout and Grusky (2018) describe this project elsewhere in this volume.
- **The Longitudinal Intergenerational Family Electronic Micro-Database (LIFE-M).** The goal of LIFE-M is to link records of births, marriages, and deaths for people born between 1880 and 1930 to construct life histories of demographic events. They will augment the vital records information by also linking to the 1880, 1900, and 1940 censuses (Bailey 2017). Beginning with birth certificates from 1881 to 1930, LIFE-M is matching the birth records to marriage records and death records. By constructing family histories across multiple generations, LIFE-M will allow study of processes of demographic change in unprecedented detail.
- **Multigenerational Longitudinal Panel (IPUMS-MLP).** IPUMS-MLP will link the 1940 census backwards to 1850, and will link the censuses to administrative records of the same period (Ruggles et al. 2017). Three kinds of linked data products are planned. First, IPUMS-MLP will produce a linked database based on the principles of the IPUMS-LRS described earlier. This version will apply new machine-learning technology to construct inter-censal links that minimize false matches and maximize representativeness. The second product will disambiguate multiple matches by using family characteristics, neighbor characteristics, and location. Preliminary work indicates that this work will raise the average linkage rate to over 50% while reducing false matches. This version of the data will underrepresent migrants and persons without consistent family members, but these problems can be mitigated through weighting. The third iteration of IPUMS-MLP will add information from administrative records. The Social Security Numident and data from LIFE-M will add information about women's maiden names and dates of birth; enlistment and draft records will add physical characteristics and dates of birth; and Social Security Death records will add dates of death. The addition of these records will further improve linkage rates and eliminate false matches.

In combination, these five resources will create a unique longitudinal database spanning the period from 1850 to the present. The investigators are working closely to make the data fully interoperable across the five projects. CLIP and AOS use shared infrastructure and will be available through FSRDCs. LIFE-M and IPUMS-MLP will be fully linked to one another, extending the power of both databases. By linking IPUMS-MLP to CLIP and to the aging surveys—hinging on the critical 1940 enumeration that they have in common—researchers will have access to data on multiple generations of ancestors.

Together, the census linking infrastructure projects will provide massive collection comprising millions of American life histories of over 170 years. This will allow the most comprehensive view of long-run changes in life-course dynamics available for any place in the world and will transform our understanding of processes of economic and social change. These data provide an opportunity to better understand geographic mobility at the national and local level; the legacy of slavery and reconstruction in America; the rise of female wage labor; and the experience of immigration from the perspective of first, second, and third generations. Perhaps most important, linked historical census data will provide insights into the great social and economic transformations of the past two centuries: the industrial revolution, massive urbanization and immigration, and the profound transitions of demographic and family behavior.

Literature Cited

Abramitzky R, Boustan LP, Eriksson K. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. Am Econ Rev. 2012; 102(5):1832–56. [PubMed: 26594052]

Abramitzky R, Boustan LP, Eriksson K. Have the poor always been less likely to migrate? Evidence from inheritance practices during the age of mass migration. J Dev Econ. 2013; 102:2–14. [PubMed: 26609192]

Abramitzky R, Boustan LP, Eriksson K. A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. J Polit Econ. 2014; 122(3):467–506. [PubMed: 26609186]

Alexander JT, Gardner T, Massey CG, O'Hara A. Work Pap. Center for Administrative Records Research and Applications, U.S. Census Bureau; Creating a longitudinal data infrastructure at the Census Bureau.

Antonie L, Inwood K, Lizotte DJ, Andrew Ross J. Tracking people over time in 19th century Canada for longitudinal analysis. Mach Learn. 2014; 95(1):129–46.

Antonie L, Inwood K, Ross JA. Population Reconstruction. Cham, Switzerland: Springer International Publishing; 2015. Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses; 217–41.

Astone NM, Mclanahan SS. Family structure, residential mobility, and school dropout: A research note. Demography. 1994; 31(4):575. [PubMed: 7890092]

Bailey MJ. The Longitudinal, Intergenerational Family Electronic Micro-Database Project. Ann Arbor, MI: Univ. Michigan; 2017. https://sites.lsa.umich.edu/life-m/

Bailey MJ, Cole C, Henderson M, Massey C. Work Pap. Dept. Econ., Univ. of Michigan; 2017. How well do automated linking methods perform in historical samples? Evidence from new ground truth.

Baskerville P, Dillon L, Inwood K, Roberts E, Ruggles S. , et al. 2014 IEEE Int Conf Big Data (Big Data). IEEE; 2014. Mining microdata: Economic opportunity and spatial mobility in Britain and the United States, 1850–1881; 5–13.

Beach B, Ferrie J, Saavedra M, Troesken W. Typhoid fever, water quality, and human capital formation. J Econ Hist. 2016; 76(1):41–75.

Bleakley H, Lin J. Thick-market effects and churning in the labor market: Evidence from US cities. J Urban Econ. 2012; 72(2–3):87–103. [PubMed: 24039316]

Bloome D, Muller C. Tenancy and African American marriage in the postbellum South. Demography. 2015; 52(5):1409–30. [PubMed: 26223562]

Blumin SM. The Urban Threshold: Growth and Change in a Nineteenth-Century American Community. Chicago: University of Chicago Press; 1976.

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

Chambers R. Regression analysis of probability-linked data. Vol. 4. Statistics; New Zealand: 2009. Official Statistics Research Series

Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011; 2(3):1–27.

Chipperfield JO, Chambers RL. Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. J Off Stat. 2015; 31(3):397–414.

Christen P. Febrl: A freely available record linkage system with a graphical user interface. HDKM '08 Proc Second Australas Work Heal Data Knowl Manag. 2008; 80:17–25.

Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. New York: Springer; 2012.

Collins WJ, Wanamaker MH. Selection and economic gains in the Great Migration of African Americans: New evidence from linked census data. Am Econ J Appl Econ. 2014; 6:220–52.

Connor DS. PhD Thesis. Univ. of Calif; Los Angeles: 2017. Putting People in their Place: Intergenerational Inequality in the Age of Mass Migration.

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

Dalzell NM, Reiter JP. Regression modeling and file matching using possibly erroneous matching variables. Work Pap. 2017 arXive: 1608.06309v3.

Dribe M, Eriksson B, Scalone F. Lund Papers in Economic Demography, No 2017. Lund Univ; 2017. Migration, marriage, and social mobility: Women in Sweden 1880–1900; 1

Eli S, Salisbury L, Shertzer A. Work Pap 22591. National Bureau of Econ. Res; 2016. Migration responses to conflict: Evidence from the border of the American Civil War.

Esslinger DR. Immigrants and the City: Ethnicity and Mobility in a Nineteenth Century Midwestern Community. Port Washington, N.Y: Kennikat Press; 1975.

Faragher JM. Sugar Creek: Life on the Illinois Prairie. New Haven: Yale University Press; 1986.

Feigenbaum J. PhD Thesis. Harvard Univ; 2016. Essays on Intergenerational Mobility and Inequality in Economic History.

Feigenbaum J. Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940. Econ J. 2017 In press.

Ferrie J, Massey C, Rothbaum J. Work Pap 22635. National Bureau of Econ. Res; 2016. Do grandparents and great-grandparents matter? Multigenerational mobility in the US, 1910–2013.

Ferrie J, Rolf K, Troesken W. Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence test scores among World War Two US Army enlistees. Econ Hum Biol. 2012; 10(1):98–111. [PubMed: 22014834]

Ferrie JP. A new sample of males linked from the public use microdata sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules. Hist Methods. 1996; 29(4):141–56.

Ferrie JP. Yankeys Now: Immigrants in the Antebellum United States, 1840–1860. New York: Oxford University Press; 1999.

Ferrie JP. Work Pap. Dept. Econ., Northwestern Univ; 2004. Longitudinal data for the analysis of mobility in the U.S., 1850–1930.

Ferrie JP. History lessons: The end of American exceptionalism? Mobility in the United States since 1850. J Econ Perspect. 2005; 19:199–215.

Goeken R, Huynh L, Lynch TA, Vick R. New methods of census record linking. Hist Methods. 2011; 44(1):7–14. [PubMed: 21566706]

Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Stat Med. 2012; 31(28):3481–93. [PubMed: 22807145]

Grusky DB, Smeeding TM, Snipp CM. A new infrastructure for monitoring social mobility in the United States. Ann Am Acad Pol Soc Sci. 2015; 657(1):63–82. [PubMed: 30111895]

Guest AM. Notes from the National Panel Study: Linkage and migration in the late nineteenth century. Hist Methods. 1987; 20(2):63–77.

Hacker JD. New Estimates of Census Coverage in the United States, 1850–1930. Soc Sci Hist. 2013; 37(1):71–101.

Hacker JD, Roberts E. The impact of kin availability, parental religiosity, and nativity on fertility differentials in the late 19th-century United States. Dem Res. 2017; 37(34):2049–1080.

Hall PAV, Dowling GR. Approximate string matching. ACM Comput Surv. 1980; 12(4):381–402.

Herscovici S. Migration and economic mobility: Wealth accumulation and occupational change among antebellum migrants and persisters. J Econ Hist. 1998; 58(4):927–56.

Hershberg T. The Philadelphia Social History Project: An introduction. Hist Methods Newsl. 1976; 9(2–3):43–58.

Hof MHP, Zwinderman AH. A mixture model for the analysis of data derived from record linkage. Stat Med. 2015; 34(1):74–92. [PubMed: 25274539]

Hopkins R. Occupational and geographical mobility in Atlanta, 1870–1890. J South Hist. 1968; 34(2): 200–213.

Hout M, Guest AM. Intergenerational occupational mobility in Great Britain and the United States since 1850: Comment. Am Econ Rev. 2013; 103(5):2021–40.

Jackson R. Index to the Seventh Census of the United States. Salt Lake City: Accelerated Indexing Systems International; 1982.

Jackson R. Index to the Eighth Census of the United States. Salt Lake City: Accelerated Indexing Systems International; 1992.

Jennings JA, Sullivan AR, Hacker JD. Intergenerational Transmission of Reproductive Behavior during the Demographic Transition. J Interdiscip Hist. 2012; 43(4):543–569.

Jensen R. Quantitative American studies: The state of the art. Am Q. 1974; 26(3):225–40.

Johnson DS, Massey C, O'Hara A. The opportunities and challenges of using administrative data linkages to evaluate mobility. Ann Am Acad Pol Soc Sci. 2015; 657(1):247–64.

Karbownik K, Wray A. Work Pap. Institute for Policy Research, Northwestern Univ; 2016a. Childhood health and long-run economic opportunity in Victorian England.

Karbownik K, Wray A. Long-Run Consequences of Exposure to Natural Disasters. 2016b. CESifo Working Paper Series. No. 6196

Katz MB. The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth-Century City. Cambridge, MA: Harvard University Press; 1975.

Knights PR. The Plain People of Boston, 1830–1860: A Study in City Growth. New York: Oxford University Press; 1971.

Lahiri P, Larsen MD. Regression analysis with linked data. J Am Stat Assoc. 2005; 100(469):222–30.

Lait AJ, Randell B. An assessment of name matching algorithms. Univ. of Newcastle Upon Tyne Comp. Sci; 1996. Technical Report Series

Laslett P, Wall R, editorsHousehold and Family in Past Times. Cambridge, England: Cambridge University Press; 1972.

Litwak E. Geographic mobility and extended family cohesion. Am Sociol Rev. 1960; 25(3):385.

Long J. Rural-urban migration and socioeconomic mobility in Victorian Britain. J Econ Hist. 2005; 65(1)

Long J, Ferrie J. The path to convergence: Intergenerational occupational mobility in Britain and the US in three eras. Econ J. 2007; 117(519):C61–71.

Long J, Ferrie J. Intergenerational occupational mobility in Britain and the US since 1850. Am Econ Rev. 2013

Malin JC. The turnover of farm population in Kansas. Kans Hist Q. 1935; 4:23–49. 164–87.

Massey CG. Creating linked historical data: An assessment of the Census Bureau's ability to assign protected identification keys to the 1960 Census. U.S. Census Bureau; 2014. CARRA Working Paper Series, 2014–12

Massey CG. Playing with matches: An assessment of accuracy in linked historical data. Hist Methods. 2017; 50(3):1–15.

Mill R. PhD Thesis. Stanford Univ; 2013. Inequality and discrimination in historical and modern labor markets.

Mill R, Stein LCD. Work Pap. Dept. of Finance, Arizona State Univ; 2016. Race, skin color, and economic outcomes in early twentieth-century America.

Modalsli J. Intergenerational mobility in Norway, 1865–2011. Scand J Econ. 2017; 119(1):34–71.

Nix E, Qian N. The fluidity of race: "passing" in the United States, 1880–1940. National Bureau of Economic Research; 2015. Work. Pap. 20828

Parman J. Childhood health and sibling outcomes: Nurture Reinforcing nature during the 1918 influenza pandemic. Explor Econ Hist. 2015a; 58:22–43.

Parman J. Childhood health and human capital: New evidence from genetic brothers in arms. J Econ Hist. 2015b; 75(1):30–64.

Parsons T, Bales RF. Family, Socialization and Interaction Process. Glencoe, Ill: Free Press; 1955.

Porter EH, Winkler WiE. Approximate string comparison and its effect on an advanced record linkage system. U.S. Census Bureau; 1997. Census Bureau Research Report, RR97/02

Prechtel-Kuskens C. The WPA census soundexing projects. Prologue Q Natl Arch Rec Adm. 2002; 34(1):72–77.

Rauscher E. Does educational equality increase mobility? Exploiting nineteenth-century U.S compulsory schooling laws. Am J Sociol. 2016; 121(6):1697–1761.

Roberts E, Ruggles S, Dillon LY, Gardarsdóttir Ó, Oldervoll J, et al. The North Atlantic Population Project: An overview. Hist Methods. 2003; 36(2):80–88.

Roesch A. Matching data using sounds-like operators and SAS® compare functions. SAS Glob Forum. 2012:122–2012.

Ruggles S. The transformation of American family structure. Am Hist Rev. 1994; 99:103–128.

Ruggles S. Reconsidering the Northwest European family system. Pop Dev Rev. 2009; 35:249–273. [PubMed: 20700477]

Ruggles S. Linking historical censuses: a new approach. Hist Comput. 2002; 1+2(publ. 2006):213–24.

Ruggles S. Intergenerational coresidence and family transitions in the United States, 1850–1880. J Marriage Fam. 2011; 73(1):136–48.

Ruggles S. Big microdata for population research. Demography. 2014; 51(1):287–97. [PubMed: 24014182]

Ruggles S, Alexander JT, Bailey MJ, Ferrie JP, Fitch CA. , et al. Building a national longitudinal research infrastructure. Univ. of Minnesota; 2017. MPC Working Paper Series, 2017–2

Ruggles S, Menard RR. The Minnesota Historical Census Projects. Hist Methods. 1995; 28(1):6–10.

Ruggles S, Roberts E, Sarkar S, Sobek M. The North Atlantic Population Project: progress and prospects. Hist Methods. 2011; 44(1):1–6. [PubMed: 22199411]

Salisbury L. Women's income and marriage markets in the United States: Evidence from the Civil War pension. J Econ Hist. 2017; 77(1):1–38.

Saperstein A, Gullickson A. A "mulatto escape hatch" in the United States? Examining evidence of racial and social mobility during the Jim Crow Era. Demography. 2013; 50(5):1921–42. [PubMed: 23606347]

Scheuren F, Winkler WE. Regression analysis of data files that are computer matched. Surv Methodol. 1993; 19(1):39–58.

Schurer K, Higgs E. Integrated Census Microdata (I-CeM); 1851–1911. 2014. U.K. Data Service. SN: 7481,

Steckel RH. Census matching and migration: a research strategy. Hist Methods. 1988; 21(2):52–60.

Stephenson C. Tracing those who left: Mobility studies and the Soundex indexes to the US Census. J Urban Hist. 1974; 1(1):73–84.

Stephenson C. Determinants of American migration: methods and models in mobility research. J Am Stud. 1975; 9(2):189–197.

Sussman MB. The Isolated Nuclear Family: Fact or Fiction? Social Problems. 1959; 6(1959):333–40.

Szołtysek M, Gruber S. Mosaic: recovering surviving census records and reconstructing the familial history of Europe. Hist Fam. 2016; 21(1):38–60.

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

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

Thorvaldsen G. An international perspective on Scandinavia's historical censuses. Scand J Hist. 2007; 32(3):237–57.

Torres C, Dillon LY. Population Reconstruction. Cham, Switzerland: Springer International Publishing; 2015. Using the Canadian Censuses of 1852 and 1881 for automatic data linkage: A case study of intergenerational social mobility; 243–61.

Turner FJ. The significance of the frontier in American history. Annu Rep Am Hist Assoc Year. 1893:199–229.

Ward Z. The not-so-hot melting pot: The persistence of outcomes for descendants of the age of mass migration. Research School of Economics, Australian National Univ; 2017. Work. Pap

Warren JR, Knies L, Haas S, Hernandez EM. The impact of childhood sickness on adult socioeconomic outcomes: Evidence from late 19th century America. Soc Sci Med. 2012; 75(8): 1531–38. [PubMed: 22809795]

Wisselgren MJ, Edvinsson S, Berggren M, Larsson M. Testing methods of record linkage on Swedish censuses. Hist Methods. 2014; 47(3):138–51.

Xie Y, Killewald A. Intergenerational occupational mobility in Great Britain and the United States since 1850: Comment. Am Econ Rev. 2013; 103(5):2003–20. [PubMed: 23970805]