HHS Public Access

Author manuscript

Hist Methods. Author manuscript; available in PMC 2019 December 20.

Published in final edited form as:

Hist Methods. 2018 ; 51(4): 246–257. doi:10.1080/01615440.2018.1507772.

Linking the 1940 U.S. Census with Modern Data

Catherine G. Massey,
Population Studies Center, Institute for Social Research, University of Michigan

Katie R. Genadek,
U.S. Census Bureau

J. Trent Alexander,
Inter-University Consortium for Political and Social Research, Institute for Social Research, University of Michigan

Todd K. Gardner, and
U.S. Census Bureau

Amy O'Hara
Stanford Institute for Economic Policy Research, Stanford University

Abstract

The U.S. Census Bureau has created a set of linkable census, survey, and administrative records that provides longitudinal data on the American population across the past eight decades. While these files include modern decennial censuses, Census Bureau surveys, and administrative records files from other federal agencies, the long time span is only possible with the addition of the complete count 1940 Census microdata. In this paper, we discuss the development of this linked data infrastructure and provide an overview of the record linkage techniques used. We primarily focus on the techniques used to produce a beta version of a linkable 1940 Census microdata file and discuss the potential to further document and extend the infrastructure.

Keywords

Census; historical demography; microdata; record linkage

1. Introduction

The U.S. Census Bureau has an established infrastructure for linking records across data sources (U.S. Census Bureau 2017). For the past several decades, the Census Bureau has developed and used probabilistic matching software to link person records in decennial census, survey, and administrative datasets. These linkages are central to the Census Bureau's mission to utilize multiple resources to evaluate survey data quality and to improve social and economic measurement. The Census Bureau's record linkage process involves appending unique and consistent linkage identifiers to all restricted internal microdata files since the late-1990s.

In order to utilize this vast amount of linkable data for research purposes, the Census Bureau initiated the Census Longitudinal Infrastructure Project (CLIP). CLIP's primary goals are to document linkable data files, expand the collection of linkable data back in time, and facilitate longitudinal research using this infrastructure. By leveraging the existing data products and linkage expertise developed within the Census Bureau, CLIP will provide an unprecedented longitudinal data resource facilitating new investigations of long-term changes in health, population, and the economy. As shown in Table 1, CLIP includes most of the Census Bureau's extensive array of data files, as well as the newly available 1940 Census data. This includes decennial censuses, Census Bureau surveys, and administrative records from across the U.S. Federal Government and states.

One of the initial goals of CLIP is to create linkages between the modern datasets and the 1940 Census microdata file, which was made available to qualified researchers via an agreement between Ancestry.com and the University of Minnesota. The 1940 Census is the first decennial census to collect income and education information for the entire population of the U.S. and also includes detailed demographic and employment information. Linkages across 1940 and contemporary data will allow entirely new analyses of an additional generation of Americans. For the first time, researchers will be able to observe both adult outcomes and detailed early-childhood neighborhood and family information for a large sample of the U.S. – enabling research on the long-term effects of early-life exposure (such as to lead pipes), neighborhood quality, and intergenerational transmissions of health, income, and education. Furthermore, the 1940 data can be linked to census and survey data from earlier time periods, extending the longitudinal analyses into multiple generations.

Appending linkage keys to the 1940 Census presented new challenges when using the Census Bureau's linking techniques because of the limited information in the historical data. While modern data often includes detailed information on respondents such as birthdates and Social Security Numbers (SSNs), the historical data only include names and census questions such as age and birthplace. Adapting the Census Bureau's existing linkage infrastructure to accommodate limited inputs required several alterations.

In this paper, we describe the current record linkage techniques used for contemporary data as well as those we used to make the beta version of the linkable 1940 Census. We describe and evaluate our methods of linking the 1940 Census to the Social Security Administration's Numident file as well as the linkages enabled between the 1940 Census and contemporary census and survey data such as the 2000 Census, 2010 Census, and the American Community Survey (ACS).

2. Linkage Techniques at the Census Bureau

The Census Bureau uses the Person Identification Validation System (PVS) to assign unique Protected Identification Keys (PIKs) to person records to facilitate deduplication and record linkage. The PVS uses a probabilistic matching algorithm (Felligi and Sunter, 1969) to compare personally-identifiable information (PII) of person records in census, survey, and administrative data to person records in a reference file constructed from the Social Security Administration (SSA) Numident file and other federal-agency administrative data such as

1040 tax returns, which provide street address. Each person in the reference file corresponds one-to-one with a unique PIK. The PVS assigns PIKs to person-records in census, survey, and administrative data whose characteristics sufficiently match the characteristics of records in the reference file. These characteristics include different combinations of Social Security Numbers (SSNs), full name, full date of birth, and address (Wagner and Layne, 2014). Research shows that when a case receives a PIK, the quality of the PIK is high. A review of this literature is in Appendix A.

Overview of the Person Identification Validation System (PVS)

The PVS follows the typical steps in record linkage: preprocessing, sorting into blocks, identifying potential matches, and resolving best matches. Person records are preprocessed to standardize the blocking and matching fields between the census file and the reference file, ensuring that similar variables align. Next, the input and reference files are sorted into "blocks" or "cuts" for comparison.

Blocking creates reasonably sized search spaces to find candidate matches, which is important with large files (Michelson and Knoblock, 2006). The SSA Numident is a cumulative file containing a record of all SSN's issued since 1936. It consists of nearly 500 million SSNs and forms the bulk of the reference file used in the matching process. The reference file is enhanced using administrative records to obtain additional variables not in the Numident, such as place of residence. The reference file also includes all transactions (e.g. name changes, corrections, etc.) associated with each SSN.[i] Consequently, the reference file is large and it is technically infeasible to compare every Numident record to every census record. To efficiently process large quantities of data, PVS processes data through modules, with each module blocking the data in different ways and comparing different fields to find matches.

PVS compares census and reference file records within several passes within a module. Each module physically breaks the data into pieces depending on whether the records agree in the blocking field and only records falling within each block are compared as potential matches. Each pass within a module employs additional blocking strategies (specified by the data processing staff) constructed from slight variations of the match fields. Subsequent passes permit more fuzziness in the match. For instance, the first pass may require that first name, last name, and year of birth match exactly, but the next pass may permit more tolerance in year of birth.

Within each pass of a module, potential matches are assigned a total score depending on the similarity of the characteristics of the input records and reference file records. The total score is the sum of the agreement and disagreement weights attributed to each matching variable (Felligi and Sunter, 1969). To assign agreement and disagreement weights for names, PVS employs a Jaro-Winkler string comparator to measure similarity between first, middle, and last names in the input and reference files (Winkler, 1995).[ii] This measure serves as a metric of how closely two names match, while allowing for some degree of

[i]Transactions occur when corrections or name changes are made on a record of a particular SSN. This allows us to observe both married and maiden names for women. The average number of transactions per SSN is 2.1 (Harris, 2014).

misspelling. For numeric variables, such as year of birth, a maximum acceptable difference between the variable value in the input and reference record is set. This also allows for creation of an interval, or band, around year of birth to permit inexact matches. Within this band, prorated agreement and disagreement weights are assigned depending on the similarity of year of birth.

Potential matches are identified within each pass of a module, and only those with an overall score greater than a user-specified cutoff score are retained as potential matches. Input records that do not receive a match in one pass move to the next pass. Once the input data has been processed through all passes of a module, potential matches are grouped into one file and sorted by person and by score.

The final step of a module evaluates the potential matches. The matches with the highest scores are processed using a decision rule to determine whether a PIK will be assigned. If one potential match has a higher score than all the other potential matches for a particular input record, then the PIK associated with that reference record is assigned to that input observation. If there are multiple potential matches for a particular input observation with the same high score, then no PIK is assigned in that module. Records that fail to find a match in a module are passed along to the next module.

PVS Modules

PVS employs a customized combination of search modules, depending on the characteristics of the input data file. The PVS modules include the Verification, GeoSearch, Movers, Name, Date of Birth (DOB), and the Household Composition modules.

The Verification Module is the first step for all data that contain SSNs. The module matches input records to the reference file by an SSN and then compares name and date of birth. If name and date of birth agree sufficiently PVS assigns the corresponding PIK to the input record.

The GeoSearch Module processes records that fail the Verification Module and is the first module for records without SSNs. This module blocks the data by the first three digits of the ZIP code (ZIP3), and then compares input and reference records falling in the same ZIP3 block. Subsequent passes within the GeoSearch module use finer definitions of geography to generate smaller groups of records to seek matches between the input record and the reference file. These passes begin with geography defined as finely as the household street address. The GeoSearch module scores potential matches by similarity in name, date of birth, and sex.

The Movers Module processes records that fail the GeoSearch module. To be eligible for the Movers Module, no member of the household can have received a PIK in any of the preceding modules and the household must consist of at least two people. This module then matches the household as a whole – considering persons living at the same address as a unit

[ii]The PVS string comparator was developed by Winkler (1995) and measures the distance between two strings on a scale from 0 to 900, where a distance score of 0 is given if there is no similarity between two text strings and a score of 900 is given for an exact match. The cutoff value for the string distance is set to 750 in the Name Search module.

and searching for matching units living together in the reference file (without regard for address).

The Name Search Module blocks records using the first letter of the first and last name fields. For instance, Alex Aron would be sorted into the A-A cut and Alex Bron would be sorted into the A-B cut, in both the census data and Numident data. The Name Search module compares input records to reference records within the same cuts. The Name Search Module scores matches using name, date of birth, and sex. The Name Search Module also accounts for instances where census records contain a nickname. For these records, the preprocessing step of the Name Search module outputs two records for these observations, one record for the nickname and one record for the formal name. For example, if the input record has the name “Bill Smith,” the formatting program will add a formal name “William” to that record. This record will then output to both the B-S cut and to the W-S cut.

The DOB Search Module blocks records using month and day of birth. Blocking on month and day of birth allows for miscoding in the year of birth. This module scores potential matches from comparisons of name, date of birth, and sex.

The Household Composition Search Module uses PIKs assigned to fellow household members to find PIKs for unmatched household members. It compares each household with at least one PIK assigned to the universe of known family members associated with that PIK observed in other data files. Within these households the Household Composition Search Module compares name, date of birth, sex, and address data to identify and score matches.

3. 1940 Census Beta PIKs

The addition of the 1940 Census greatly enhances the capabilities of CLIP for researching long-term outcomes of early-life circumstances and events as well as intergenerational transmissions of income, education, and health. To ensure we assigned the PIKs reliably and accurately, we substantially altered PVS to produce a beta version of 1940 Census with PIKs.

Datasets processed through PVS ideally have a core set of PII including name, date of birth, place of residence, and SSN (when available). PVS was designed to link recent Census Bureau surveys to each other and to administrative records, thus the PVS process required significant modifications to work effectively with the limited PII in historical census data. Using PVS software and modules as its core, we developed the Adaptable Record Matching System (ARMS) to assign PIKs to 1940 Census respondents.

Linking Variables in the Adaptable Record Linkage System

The ARMS is designed to assign PIKs to the 1940 Census using name, age, state or country of birth, state of residence in 1940, state of residence in 1935, and parents’ names. The PVS process has well-established techniques for matching on name and date of birth, but new techniques were developed to make use of the additional matching variables.

ARMS processes individual names using the same methods as PVS system. The text-string comparator built into PVS compares first and last names. Middle initials and middle names

are available for a small number of records and are also compared using the text string comparator in PVS.

Since date of birth is not available in the 1940 Census, ARMS matches records using age. The reference file contains full date of birth information, which is used to calculate age on April 1, 1940. The resulting variable “age on April 1, 1940” is comparable with the age variable collected in the 1940 Census (but provides considerably less detail than the date of birth information typically used in PVS).

Geographic variables used in ARMS include birthplace and location in 1940 and 1935. Although the 1940 Census contains as detailed geography as street address, there is no comparable information available in the reference file.[iii] In order to use state or country of birth, Census Bureau staff harmonized the place of birth information in the reference file and the 1940 Census birthplace variable. For SSNs issued within a year of 1940, ARMS compares the state of residence reported in the 1940 Census to the state of residence inferred from the first three digits of the SSN. For SSNs issued within a year of 1935, the SSN-inferable state is compared to the “place of residence in 1935” variable that was collected on the 1940 Census.

Finally, ARMS uses parents’ names as linking keys when available on both the 1940 Census and the reference file. In the 1940 Census, parents’ names can be inferred between co-resident people by using “relationship to household head” responses. Parents’ names are also available in the Numident. We appended parents’ names not only for youth, but also for any parent-child relationship that is evident from 1940 Census relationship values (such as spouse-stepchild, parent-grandparent, etc.). Parents’ names serve as a means to distinguish between two similar potential matches but are never required to establish a match, nor do they prohibit a match if the parents’ names on the census and reference file do not match.

Operationalizing New Linking Variables into ARMS

The ARMS incorporates the use of additional variables by developing new modules for birthplace (Birthplace Module), age on April 1, 1940 (Age Module), state of residence in 1940 (Geo1940 Module), and state of residence in 1935 (Geo1935 Module). The scored matching fields include first name, middle initial or middle name, last name, age, sex, state or country of birth, and mother’s and father’s first name. The parameter file allows for up to a maximum of a two-year age difference between a census record and a reference file record.

Figure 2 illustrates how records were processed through the modules. ARMS first processes data through the new Birthplace Module, which blocks data by state or country of birth. Next, ARMS processes unmatched records through the Name Search Module, which operates identically to the Name Search Module used in the Production PVS, scoring potential matches using the match fields available in the 1940 Census. Following the Name

[iii] In modern linking conducted at the Census Bureau, linkages to corresponding tax year provide addresses in the reference file (e.g., 1999 tax year returns are used to PIK the 2000 Census). We are unaware of any digitized historical administrative data that would provide a reference point for addresses in the 1940 Census.

Search Module, records failing to find a match are compared by the Age Module. This module creates blocks by age on April 1, 1940.

For records that have still not received a PIK, ARMS processes them through the new Geo1940 Module. This module compares records within state of residence reported in the 1940 Census and the state where SSNs were issued, including only records from the reference file where SSNs were known to have been issued within one year of 1940. This module blocks the reference file by the state a person's SSN was issued and blocks the census by state of residence in 1940.

The Geo1935 Module is identical to the Geo1940 Module except that it blocks on state of residence in 1935 and matches age in 1935 to age of SSN issue, allowing a one-year interval around age. This module effectively considers only records that still not assigned a PIK and obtained an SSN in 1936 (the first year that SSNs were issued). Again, the age difference between a census and reference record must two years or less.

The last attempt to pick up additional matches occurs in the County of Birth module for 1940 respondents under the age of 2. This module assumes infants are unlikely to move between birth and the 1940 Census, allowing us to observe their likely county of birth in 1940. This is compared to city of birth aggregated up to the county level in the Numident to provide additional detail to distinguish between potential matches.[iv]

ARMS Reference File

The CLIP historical linkages rely on the quality and coverage of the reference file, which is largely comprised of Numident data. Furthermore, we can only assign PIKs to individuals alive in 1940 who eventually received a SSN, which depends on whether they entered the work force, their industry, and other factors such as mortality and age. When first introduced, not all individuals received SSNs (including self-employed workers and those working in agriculture, domestic service, labor, and government), and only individuals under the age of 64 were considered (Puckett 2009). Eventually, anyone who applied received an SSN (Social Security Board 1938), but timing is an important factor and not everyone had a reason to apply. For instance, we cannot assign a PIK to anyone who was alive in 1940 and died prior to receiving a SSN. Also, we cannot assign a PIK to women who were enumerated with their birth (or maiden) name in 1940 and did not file their birth name with the SSA.

Several Numident variables are particularly critical for ARMS, including date of birth (used to determine age on April 1, 1940), state where SSN was obtained (for the Geo1940 and Geo1935 Modules), and names (name is a core matching variable, and parents' names are used to choose among multiple potential matches and to enhance the reference file with potential maiden names).

[iv]We use Black et al.'s (2015) GNIS codes for city of birth listed in the Numident. We then aggregate city up to the county level for the match.

Date and place of birth.—The vast majority of people in the Numident born before 1941 have a complete date of birth. Of those born before 1941, approximately 0.55% are missing year of birth. Only 0.25% of records that have year of birth are missing month of birth and 0.56% are missing day (0.24% are missing both month and day). The reported dates of birth that show a regular distribution (for instance, there is not a disproportionate number of cases with a date of birth of January 1st), which may suggest minimal age heaping or date heaping. The Numident also has excellent availability of place of birth, containing place of birth for more than 97% of records born on or before April 1, 1940.

State where SSN was obtained.—For SSNs issued before 1972, the first three digits reflect the location of the SSA office that issued the SSN. For the subset of the population who obtained an SSN in 1936 or 1939–1941, the state of SSN issuance is compared to state of residence in 1935 and 1940, as reported in the 1940 Census. The late 1930s and early 1940s were a peak in SSN issuance for those born in the 1910s and 1920s. Figure 1 charts age at SSN issuance by birth cohort using Social Security Numbers in the publicly-available 2011 Social Security Death Index. As Figure 1 shows, younger cohorts often obtained SSNs at ages when they would be entering the labor force, ages 15 to 20. Working-age adults in qualified occupations would have largely received their SSN in 1936, when SSNs were first issued. This explains the patterns we see for those born before 1920. Using the Geo1935 and Geo1940 modules, these states of issuance could be compared to information from the 1940 Census.

Parents' names.—The Numident contains parents' first and last names, and ARMS used these variables for matching. As Table 2 shows, at least one parent's name is recorded in the Numident for 97% of those aged 0–9 years in 1940, and for 92% of those aged 10–19 in 1940. These rates decline for those aged 20 and up. Parents' names are available in the 1940 Census at similar rates for those aged 0–19, though these rates decline even more sharply for those aged 20 and up. In the census data, we only observe parents' names when the parent and child are co-resident. Due to these coverage issues, ARMS's use of parents' names is most successful for those who were under age 20 in 1940.

The ARMS also uses parental names to infer maiden names for any record missing maiden name in the Numident. Many women in the 1940 Census were enumerated under what would later become their maiden or birth name. Women enumerated under a maiden name in the 1940 Census will receive an accurate PIK only if their maiden name is in the reference file. For women living under their maiden name in 1940 and who married before they worked and obtained an SSN, the reference file may not contain a record of that woman under her maiden name, and an accurate link to the 1940 Census will not be made.

Women's surname changes at marriage are a common problem for record linkage projects. For instance, IPUMS's Linked Representative Samples (IPUMS-LRS) focused on men, married couples, and women whose marital status did not change over time. To adjust for this limitation, ARMS uses an expanded reference file specifically designed to improve PIK assignment for women. When a woman has a different last name from her father in the Numident (and had thus presumably changed her name in marriage), ARMS reference file

includes a synthetic “maiden name” record for the woman, with her father’s surname in the last name field.

4. Progress on the 1940 Census and Next Steps

The Census Bureau acquired the digitized and coded 1940 Census microdata from the IPUMS. Census staff have completed most pre-processing for name and birthplace and preliminary PIK assignment is underway, with a completed beta version of the 1940 PIKs now available for pilot CLIP projects.

Our main goal with production of the beta 1940 PIKs is to produce a high-quality file suitable for inference and research. Given the abundance of survey methodologies designed to cope with sampling and representativeness issues, we chose to focus on minimizing instances of Type I error (false links) instead of trying to minimize both Type I error and Type II error (false nonlinks). Given this focus, we examine the beta 1940 PIKs along several dimensions including accuracy, PIK rates, and representativeness.

Beta 1940 PIK Accuracy

To test the proposed techniques to assign PIKs to the 1940 Census, we simulate the 1940 ARMS/PVS process using 2000 Census long-form data. The 2000 data contain full name, age, place of birth, and PIKs assigned by PVS. Using 2000 Census data allows us to compare the PIKs assigned by the formal PVS to those assigned by ARMS. Although imperfect, PVS results for the 2000 long-form provide a “truth deck” of sorts, since PVS assigned PIKs to the 2000 Census using detailed PII unavailable in the 1940 Census, including exact date of birth and street address. We simulate ARMS on these records by “hiding” this additional detail and use only linkage variables available in the 1940 Census.[v] By comparing results from PVS and ARMS, we can calibrate ARMS parameters to increase the accuracy of the 1940 process.

The blocking schemes, scoring, and cutoff values used in ARMS for the beta 1940 Census PIKs were chosen using the 2000 Census “truth deck” and multiple simulations aimed at minimizing instances of incorrect links. We were able to find scoring and cutoff values that limited the instance of Type I error in our simulations to 3.9% for PIKs assigned to the 2000 Census using ARMS. We then applied ARMS calibrated to these scoring and cutoff values to the 1940 Census to produce the beta PIKs. A simulated error rate of 3.9% may suggest the beta 1940 PIKs are highly accurate, but we also implicitly test accuracy of the 1940 beta PIKs using linkages of the 1940 Census to the 2000 Census by PIK.

As an independent check of the accuracy of the 1940 beta PIKs, we link the 1940 Census to the 2000 Census by PIK and compare state of birth. Although state of birth is used to assign PIKs to the respondents in the 1940 Census, it was not used as a linkage key in the assignment of PIKs in the 2000 Census. Therefore, state of birth agreement in the 1940 and 2000 sample linked by PIK allows for an implicit test of accuracy in the 1940 PIKs for

[v]Although address stings are available for many records in the 1940 Census, we do not observe addresses for observations in the reference file. Consequently, 1940 respondent address is not useful for PIK assignment.

respondents ages 0–30 in 1940 (who are likely to survive the 60 year gap between 1940 and 2000). We find state of birth agrees for 94.0% of cases ages 0–9 in 1940, 93.8% of cases ages 10–19, and 93.0% of cases ages 20–29. These rates are slightly lower than some rates found in the historical linkage literature, including the 2% (Goeken et al. 2011) to 4% (Bailey et al. 2017a) error rate estimated for IPUMS-LRS.

Beta 1940 PIK Rates

There exists a tradeoff between linkage rates and accuracy. If not concerned about accuracy, we could assign PIKs to 100% of respondents in the 1940 Census. Instead, we focus on minimizing linkage error. Despite this choice, we still achieve match rates that are relatively high for historical linkage, which range anywhere from 7% (IPUMS-LRS 1870–1880 sample) to 46% (Bleakley and Ferrie 2016).

Figure 3 displays the number of respondents with PIKs by age and sex for the 1940 Census. Overall, 40.7% of 1940 respondents receive a PIK. Strikingly, PIK rates vary little between men and women, but there is a clear drop off of PIK assignment for older cohorts. PIK rates are high for those 20 years and younger (ranging from 69.4 to 76.6 percent), quickly fall to 34.4% for those 30–34 years old, and are as low as 4.7% for those 70 years and older. Higher match rates for younger cohorts result from using parents' names as additional linkage keys, while lower match rates in the older cohorts are a function of both the unavailability of additional linkage keys (specifically parents' names) and the rollout timing of SSNs. When first introduced in 1936, not every type of worker was required to obtain an SSN (including self-employed workers such as farmers) and it was not until the 1950s and 1960s that the issuance of an SSN became population wide (Puckett 2009). We are unable to assign a PIK to any individual who never obtained an SSN, and thus limited in our ability to PIK older cohorts. Given the focus of CLIP on linking individuals forward in time from 1940 to 60 or more years later, the lower PIK rates for older cohorts should not be problematic.

Beta 1940 PIK Representativeness

As the PIK rate increases, so does the representativeness of the linked data. To formally test the representativeness of the beta 1940 PIKs, we regress an indicator variable for whether an observation received a PIK on various household and individual characteristics in Table 3. Ideally, each coefficient would be statistically insignificant or, in the case of large samples, a precisely estimated, very small number, indicating no difference in the mean observable characteristics of the matched and unmatched samples. Many of the coefficients are statistically significant because of the large sample size, but we find little difference between matched and unmatched cases in terms of farm status, urban status, family size, sex, age, and education attainment while estimating by age group. However, we are much more likely to assign PIKs to white respondents and native-born respondents. The lower PIK rates of foreign-born respondents may result in part from high return migration rates and the necessity that migrants remain in the US long enough to enter the workforce and receive an SSN. The finding of lower match rates for non-white respondents aligns with previous findings of lower PIK rates for minorities (Bond et al. 2014).

We can also assess representativeness relative to the pool of individuals alive in 1940 in the Numident. Table 3 shows that PIK rates relative to the Numident are higher for younger individuals, whites, and those born in the US. Interestingly, PIK rates are much higher for those ages 65 and older than in Figure 3, suggesting the lower PIK rates relative to the 1940 Census are more a function of who had a SSN than of the matching procedure. Similar to the representativeness tests relative to the 1940 Census, we do find significant differences in some characteristics between the matched and unmatched observations in Table 3, which suggests the sample with PIKs is unrepresentative of the pool of individuals alive in 1940 who received an SSN and are in the Numident. However, unrepresentativeness of linked samples is a common problem in historical linking and is easily addressable using inverse propensity score reweighting (Bailey et al. 2017a, Bailey et al. 2017b).

Next Steps

We are currently working to further assess the representativeness of the 1940 Census PIKs, as well as the representativeness of the 1940–2000 linked data. We are also developing basic metadata for the linkable data. Technical documentation, such as how place of birth is coded, the use of alternate names, birthplace harmonization, and the reliability of SSN area numbers, is also required. To improve the linkages of the 1940 Census, we will further examine how to improve accuracy by examining incorrect linkages (identified by discordant information when linked to other files by PIK), and how to improve PIK rates through use of additional historical administrative data such as the World War II enlistment records.

5. Capabilities of CLIP

The Census Longitudinal Infrastructure Project will create an unparalleled source of uniformly processed linkable census, survey, and administrative data. PIKs will facilitate linkage across files, enabling research on population dynamics, migration, life course, social mobility, generational linkages, and socio-economic status.

The 1940 Census plays a pivotal role in CLIP by allowing linkages to the ACS, 2010 Census and 2000 Census, as well as survey and administrative records. Table 4 reports the PIK rates of the core files in CLIP. Each of the contemporary core files exhibit high PIK rates, providing millions of observations linkable across time and generations. The intersections of these core files are reported in Table 5. Any two core files provide longitudinal data surpassing any of today's currently available longitudinal survey data in terms of size and scope. Such large datasets allow for broader analyses of population change and are more likely to include underrepresented populations.

In addition to the current data available through CLIP, the ARMS process used to assign PIKs to the 1940 Census data provides us with a tool to assign PIKs to other historical data. For instance, we are currently working on adding historical Current Population Survey (CPS) data, and we are developing methods to digitize and assign PIKs to the 1950–1990 Censuses. To differing degrees, all of these files will need to use ARMS processes, since SSNs and residential addresses are rarely available in these files. For the pre-1940 records, we may not be able to use ARMS at all—since the heart of ARMS links records to the Social Security Administration's Numident file, and respondents will age out of the

Numident in earlier censuses. However, we will be able to take advantage of external efforts to link older data, such as the 1850–1930 census microdata and enlistment records, to the 1940 data,[vi] which we can simply merge together using identification numbers in the 1940 Census. The 1940 census, then, plays a pivotal role in allowing long-range linkages spanning more than a century.

The beta version of the CLIP infrastructure is currently in use by nine pilot research projects. The Census Bureau is continuing to improve and document the linkable data files, and to establish procedures for external researchers to gain approval to access the files in the Federal Statistical Research Data Centers. Once documented and piloted, CLIP will provide an unparalleled resource for population studies of the United States in the twentieth century.

Appendix A:: PIK Quality

Research shows that when a case receives a PIK, the quality of the PIK is high. Using Medicare Enrollment Database (MEDB), Indian Health Service (IHS) patient registration files, and commercial data, Layne et al. (2014) assess the error rates of several modules. For each dataset, they compare a deterministically matched PIK from the administrative record SSN to a PIK from PVS probabilistic matching to assess the PIK error rate. The MEDB data has a PIK error rate ranging from 0.005 percent to 1.174 percent depending on the module (Layne et al., 2014). The IHS data has a slightly higher PIK error rate (e.g., 0.050 percent in the GeoSearch module), and the commercial data has a PIK error rate ranging from 0.185 percent (GeoSearch) to 4.177 percent (Name Search) (Layne et al., 2014). Since this analysis processed all records through each module (i.e., records were not removed once they received a PIK), these estimates should not be taken as PVS error rates. The actual error rate for PVS would depend on what proportion of records cascaded into each module, which would be a function of the characteristics of the input data. Research is underway to estimate error rates for the entire PVS process.

Although cases with PIKs are likely to be identified reliably, there is evidence that certain types of records are systematically less likely to receive PIKs than others. In an analysis using 2001 ACS and the 2002–2005 CPS Annual Social and Economic Supplement (ASEC), Meyer and Goerge (2011) find statistically significant differences in PIK rates by race, household size, citizenship, and rural status. Using 2009 ACS data, Mulrow et al. (2011) find substantial geographic differences in PIK assignment. Bond et al. (2014) find the assignment of PIKs by the formal PVS process to be non-random for 2009 and 2010 ACS data. They show migrants, those without health insurance, and those in poverty are less likely to have PIKs, while those in the military and the highly educated are more likely to

[vi]Outside of the Census Bureau, many scholars link person records across public (pre-1950) decennial census data (for example, see Ferrie, 1996; Collins and Wanamaker, 2014; Long and Ferrie, 2013; and Abramitzky et al. 2014). At the forefront, the IPUMS project produced the most comprehensive linkages of individuals across the 1850–1930 U.S. censuses with their IPUMS Linked Representative Samples (Ruggles, 2006 and 2011; Goeken et al. 2011). These linkages began with individuals in the 1880 complete-count census file linked to the other earlier censuses to produce seven pairs of linked samples. The IPUMS is expanding their historical linkages to the full-count historical population censuses to construct the Multigenerational Longitudinal Panel (IPUMS-MLP). A collaborator of IPUMS-MLP, the Longitudinal, Intergenerational Family Electronic Micro-database (LIFE-M) project, uses historical vital records to improve the linkages of women and minorities across the complete-count1880–1940 censuses (Bailey et al. 2017). Similarly, the Early Indicator's project link Union Army enlistees and their families across numerous decennial censuses (Costa et al. 2017).

receive a PIK. The most obvious source of bias in the PIK process is that the reference file is built from federal agency records. Anyone who does not have an SSN or Individual Taxpayer Identification Number (ITIN) will necessarily not receive a PIK. The better a case is represented in federal agency records, the more likely that case is to receive a PIK.

Based on research on bias in the PIK process, the Census Bureau has implemented improvements to PVS, including the development of new matching modules and the incorporation of additional records into the reference file. Bond et al. (2014) find these enhancements led to higher validation rates for the 2010 ACS. Groups that experienced the highest increase in PIK assignment as a result of these improvements are those ages 0–2, non-U.S. citizens, the uninsured, those with no schooling completed, and those living in small multi-unit buildings.

References

Bleakley H and Ferrie J (2016). “Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital across Generations.” Quarterly Journal of Economics 131(3): 1455–1495. [PubMed: 28529385]

Bailey MJ, Cole C, Henderson M and Massey CG (2017). “How Do Automated Linking Methods Perform? Evidence from the Life-M Project.” Retrieved September 21, 2017. Available at http://www-personal.umich.edu/~baileymj/Bailey_Cole_Henderson_Massey.pdf.

Bailey M, Cole C and Massey CG (2017). “Representativeness and False Links in the 1850–1930 IPUMS Linked Representative Historical Samples.” Retrieved October 1, 2017. Available at http://www-personal.umich.edu/~baileymj/Bailey_Cole_Massey.pdf.

Bond B, Brown J, Luque A & O’Hara A, 2014 The Nature of Bias When Studying Only Linked Person Records: Evidence from the American Community Survey. CARRA Working Paper #2013–08.

Collins WJ & Wanamaker MH, 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), pp. 220–52.

Corson JJ, 1938 Administering Old-Age Insurance. Social Security Bulletin, 1(5), pp. 3–6.

Costa DL, DeSomer H, Hanss E, Roudiez C, Wilson SE and Yetter N (2017).Union Army Veterans, All GrownUp. Historical Methods 50(2): 79–95. [PubMed: 28690347]

Fellegi IP & Sunter AB, 1969 A Theory for Record Linkage. Journal of the American Statistical Association, Volume 64, pp. 1183–1210.

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

Goeken R, Huynh L, Lynch TA & Vick R, 2011 New Methods of Census Record Linking. Historical Methods, 44(1), pp. 7–14. [PubMed: 21566706]

Grusky DB, Smeeding TM & Snipp CM, 2015 A New Infrastructure for Monitoring Social Mobility in the United States. Annals of the American Academy of Political and Social Science, 657(1), pp. 63–82. [PubMed: 30111895]

Harris B, 2014 Transgender Labor Supply, Employment, and Earnings Gaps: Evidence from the Federal Administrative Records and the American Community Survey. CARRA Working Paper.

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

Massey CG, 2014a Creating Linked Historical Data: An Assessment of the Census Bureau’s Ability to Assign Protected Identification Keys to the 1960 Census. CARRA Working Paper 2014–12.

Massey CG, 2014b Playing with Matches: An Assessment of Match Accuracy in Linked Historical Data. CARRA Working Paper 2014-XX.

McGaughey A, 1994 The 1995 Bureau of the Census Computer Name Standardizer Documentation. Statistical Research Division Research Paper.

Meyer Bruce D., and Goerge Robert. 2011 “Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation,” Center for Economic Studies (CES) Working Paper Series, U.S. Census Bureau.

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

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

NORC, 2013 PVS Research: Task 4, Further PVS Research Final Research Report.

Puckett C. 2009 The story of the social security number. Social Security Bulletin 69 (2):55–74 [PubMed: 19697506]

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

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

Ruggles S, Alexander JT, Genadek K, Goeken R, Schroeder MB, and Sobek M. Integrated Public Use Microdata Series: Version 5.0. Minneapolis: University of Minnesota.

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

Social Security Death Master File courtesy of SSDMF.INFO.

Social Security Board, Bureau of Old-Age Insurance, Analysis Division. 1938 Old-age insurance. Social Security Bulletin. 1(5): 49–54

Census Bureau U.S., 2017 “Data Linkage Infrastructure.” Webpage at https://www.census.gov/about/adrm/linkage.html. Accessed July 31, 2017.

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 WE, 1995 Matching and Record Linkage In: Business Survey Methods. New York: J. Wiley, pp. 355–384.

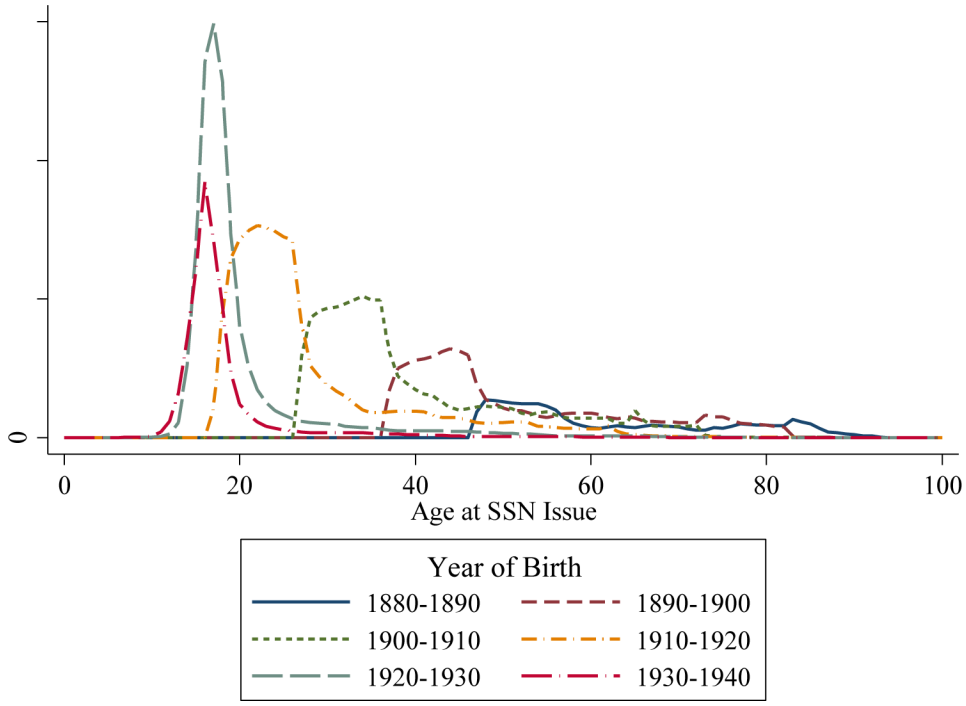

Figure 1. Age at Time of SSN Issue by Birth Cohort
Source: 2011 Social Security Death Index

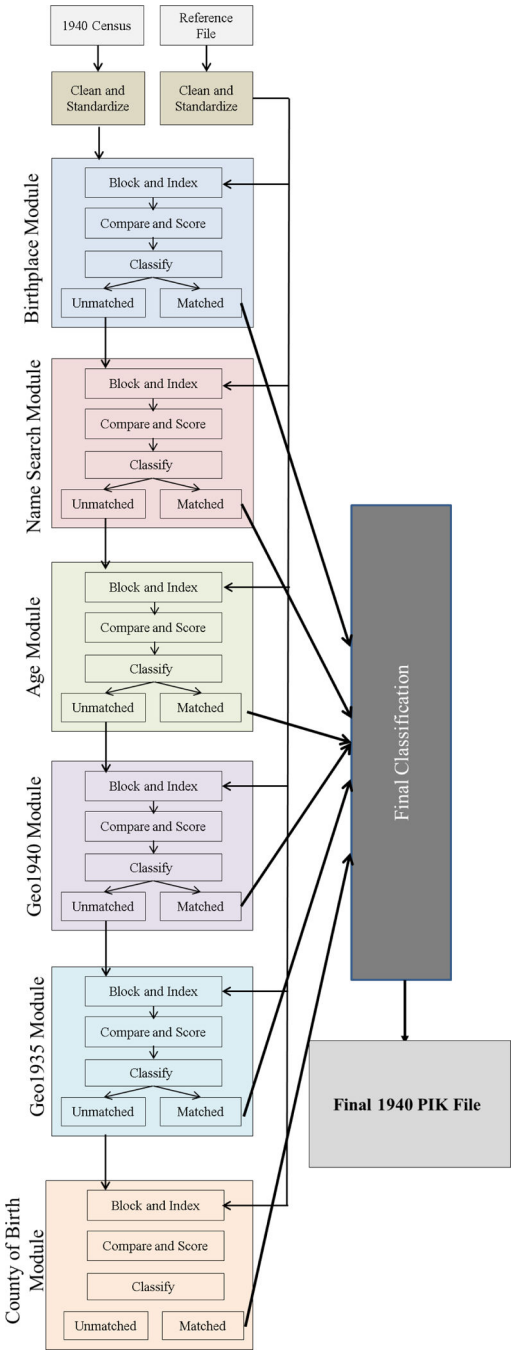

Figure 2. Adaptable Record Linkage System Flow

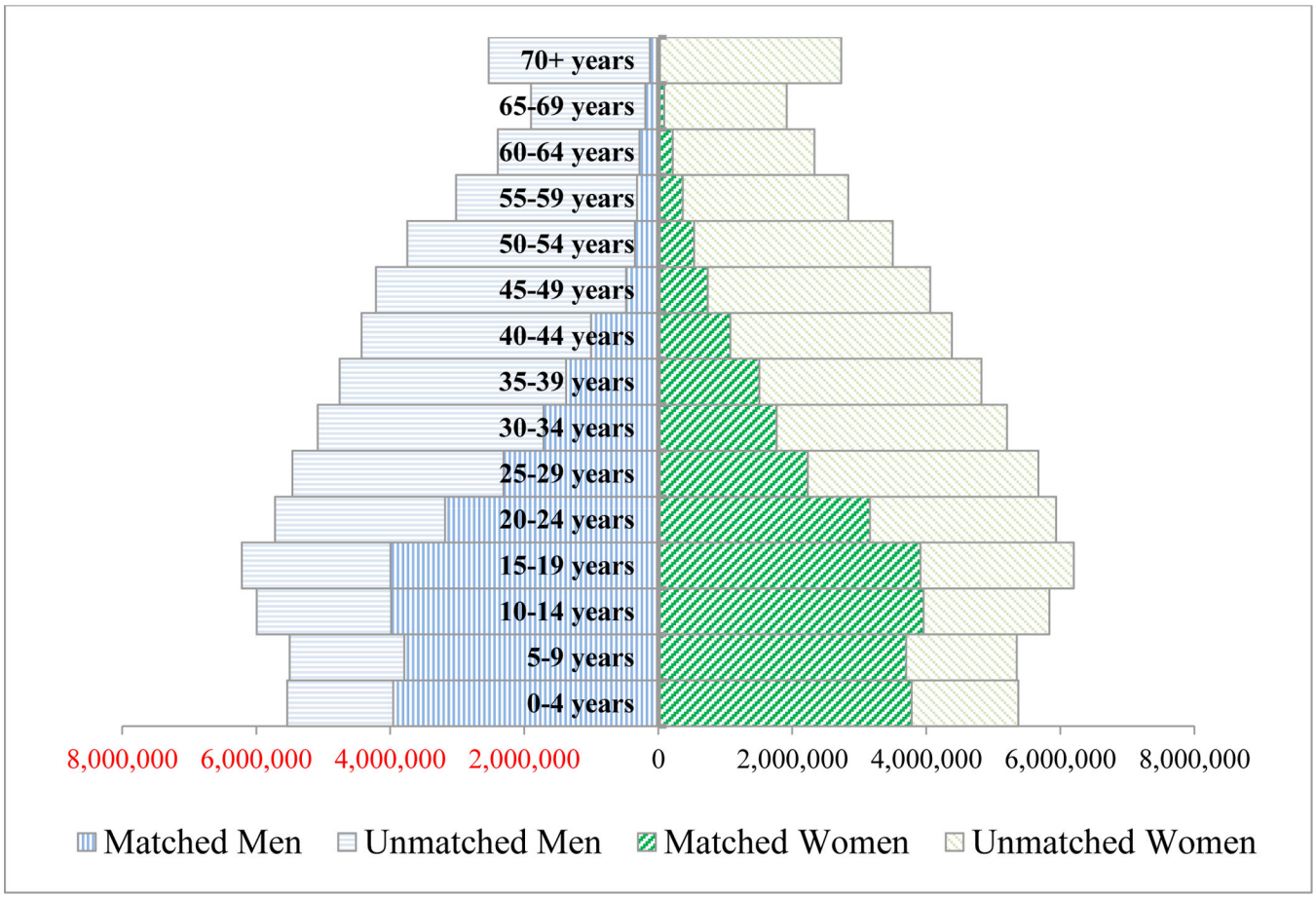

Figure 3. Beta 1940 Census PIK Rate by Cohort and Sex. Source: 1940 Census.

Table 1.

Major CLIP Datasets, linkable and available at the Census Bureau

Census Bureau Data
Decennial Census, 1940, 2000, and 2010
American Community Survey, 2001–2016
Current Population Survey (March), 1973, 1978, and 1981–2016
Survey of Income and Program Participation, 1984–2014 American Housing Survey, 2013 and 2015
Administrative Data from other Federal Agencies
Social Security Administration Numident File, 1936-present
Social Security Administration Death Master File, 1978-present
Indian Health Service Patient Registration File, 1999-present
Public and Indian Housing Information (Public Housing residents), 1999-present
Tenant Rental Assistance Certification (Section 8 participants), 1999-present
Federal Housing Administration Loan Information, 2000-present
Selective Service System Registration, 1999-present
U. S. Postal Service Change of Address File, 2010-present

Table 2.

At Least One Parent's Name is Available in the Numident and the 1940 Census

	Numident		1940 Census	
Age in 1940	**Observations**	**% of Total Numident Population**	**Observations**	**% of Total 1940 Population**
0–9	27,420,064	97	20,791,049	96
10–19	27,897,222	92	21,531,326	89
20–29	19,662,524	70	8,608,777	38
30–64	15,241,391	28	4,624,241	8
65+	1,120,056	47	36,823	0
Total	91,339,664	63	55,592,216	42

Source: SSA Numident data and the 1940 Census.

Table 3.

Representativeness of Beta 1940 PIKs relative to the Numident

	Full Sample PIK Rates	Full-Sample Regression Coefficients	US-Born PIK Rates	US-Born Regression Coefficients
Age 0–9	54.56%	0.1323	64.40%	0.2230
		(0.0003)		(0.0004)
Age 10–19	53.00%	0.0969	58.48%	0.1714
		(0.0003)		(0.0004)
Age 20–29	40.45%	0.0346	42.44%	0.0941
		(0.0003)		(0.0004)
Age 30–64	25.82%	0.0069	26.26%	0.0464
		(0.0003)		(0.0003)
Age 65+	29.44%	(reference)	30.35%	(reference)
White	48.29%	0.1816	52.43%	0.1780
		(0.0002)		(0.0003)
Black	33.18%	0.0005	35.13%	−0.0113
		(0.0002)		(0.0004)
Missing Race	20.00%	−0.1603	20.47%	−0.1546
		(0.0002)		(0.0003)
Other Race	10.08%	(reference)	21.34%	(reference)
New England Division	42.79%	0.2045	42.79%	0.1100
		(0.0002)		(0.0002)
Middle Atlantic Division	40.28%	0.1798	40.28%	0.0857
		(0.0001)		(0.0002)
East North Central Division	42.36%	0.2056	42.36%	0.1111
		(0.0001)		(0.0002)
West North Central Division	46.18%	0.2231	46.18%	0.1285
		(0.0001)		(0.0002)
South Atlantic Division	36.65%	0.1479	36.65%	0.0527
		(0.0001)		(0.0002)
East South Central Division	36.09%	0.1451	38.13%	0.0493
		(0.0001)		(0.0002)
West South Central Division	31.47%	0.1343	37.71%	0.0470
		(0.0001)		(0.0002)
Mountain Division	21.04%	0.1572	33.53%	−0.0029
		(0.0001)		(0.0003)
Pacific Division	18.70%	0.0813	27.05%	(reference)
		(0.0000)		
Foreign Born	7.25%	−0.3768		–

	Full Sample PIK Rates	Full-Sample Regression Coefficients	US-Born PIK Rates	US-Born Regression Coefficients
		(0.0001)		–
Intercept		0.1953		0.2306
		(0.0004)		(0.0005)
R-Square		0.2204		0.1813
Observations		138,592,834		124,970,720

Notes: This table reports PIK rate by demographic category as well as coefficients from a regression of y=1 for linked records, y=0 for unlinked records in the Numident file. Standard errors in parentheses.

Source: Social Security Administration Numident File

Table 4.

PIK Rates of Core CLIP Files

Data Source	Unique PIKs in Year	Total Observations	% PIKed
2001–2013 ACS	44,301,421	48,570,815	91.2%
2010 Census	273,041,579	308,745,538	88.4%
2000 Census	247,117,007	281,421,906	87.8%
1940 Census	53,783,480	132,032,406	40.7%

Note: Only unique respondents across 2001–2013 ACS samples included here. Source: 2001–2013 American Community Survey, 2010 Census, 2000 Census, and 1940 Census

Table 5.

Intersection of Core CLIP Files

	2001–2013 ACS	2010 Census	2000 Census Short-form	2000 Census Long-form	1940 Census
2001–2013 ACS	41,520,665	-	-	-	-
2010 Census	37,203,805	273,041,579	-	-	-
2000 Census Short-form	33,182,329	201,070,146	247,117,007	-	-
2000 Census Long-form	6,349,514	32,678,878	40,055,782	40,055,782	-
1940 Census	3,243,687	15,303,883	24,556,827	4,288,958	53,783,480

Source: 2001–2013 American Community Survey, 2010 Census, 2000 Census, and 1940 Census