Creating a Longitudinal Data Infrastructure at the Census Bureau

J. Trent Alexander
Center for Economic Studies
U.S. Census Bureau
J.Trent.Alexander@Census.gov

Todd Gardner
Center for Economic Studies
U.S. Census Bureau
Todd.Gardner@Census.gov

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

Amy O'Hara
Center for Administrative Records Research & Applications
U.S. Census Bureau
Amy.Ohara@Census.gov

Abstract

In-depth analyses of population dynamics require longitudinal data spanning multiple decades. The Census Bureau has initiated a project to create a core set of linkable census, survey, and administrative records that would provide data on the American population across seven decades. The core linkable data files consist of the 1940 Census, the 2000 Census, the 2010 Census, the American Community Survey, and the Current Population Survey. Using the core as a foundation, researchers can then integrate administrative records or other external data sources, depending on their research objectives. In this paper, we discuss the development of this project and provide an overview of the record linkage techniques that enable the creation of longitudinal data of this magnitude. We also report our progress on building a linkable version of the 1940 Census.

1. Introduction

The Census Bureau has initiated a project to create an integrated set of linkable data files from the decennial Censuses, surveys, and administrative records files. The Core Longitudinal Infrastructure Population Project (CLIPP) leverages existing longitudinal products and linkage expertise developed within Census Bureau and by academic researchers. Once complete, CLIPP will provide an unprecedented longitudinal data resource facilitating new investigations of long-term changes in health, population, and the economy.

The CLIPP project builds on the Census Bureau's established infrastructure for linking records across data sources. 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 most comprehensive internal efforts at record linkage involve appending unique and consistent linkage identifiers (referred to as Protected Identification Keys, or PIKs) to all restricted internal microdata files since the late-1990s. CLIPP will include most of Census Bureau's extensive array of PIKed data files.

The CLIPP team benefits from close collaboration with academic researchers, particularly those with interests in historical demography, economic history, and demographic record linkage. The National Research Council's Committee on National Statistics is also supporting the creation of a long-term Census-based infrastructure to facilitate the study of social and economic mobility, known as the American Opportunity Study (AOS) (Grusky, Smeeding, and Snipp, 2015). The AOS team will leverage the CLIPP infrastructure and is working closely with CLIPP staff on developing and documenting best techniques for record linkage and access protocols.

The CLIPP team is also collaborating with the Minnesota Population Center (MPC). MPC's microdata infrastructure and linkage projects provide a good model for documentation, dissemination, and historical record linkage techniques (Ruggles, 2006; Ruggles et al., 2010; Goeken, Huynh, Lynch, and Vick, 2011). MPC provided CLIPP with the complete 1940 Census data and has been involved with CLIPP since its initial planning.

One of the initial goals of CLIPP is to append PIKs to the complete 1940 Census microdata file, which was recently made available to qualified researchers via an agreement between Ancestry.com and the University of Minnesota. Once assigned PIKs, person records in the 1940 Census will be linkable to other data assigned PIKs, such as the 2000 Census, the American Community Survey (ACS), the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), administrative records, and ultimately other external data sources.

Figure 1: Data Infrastructure for CLIPP

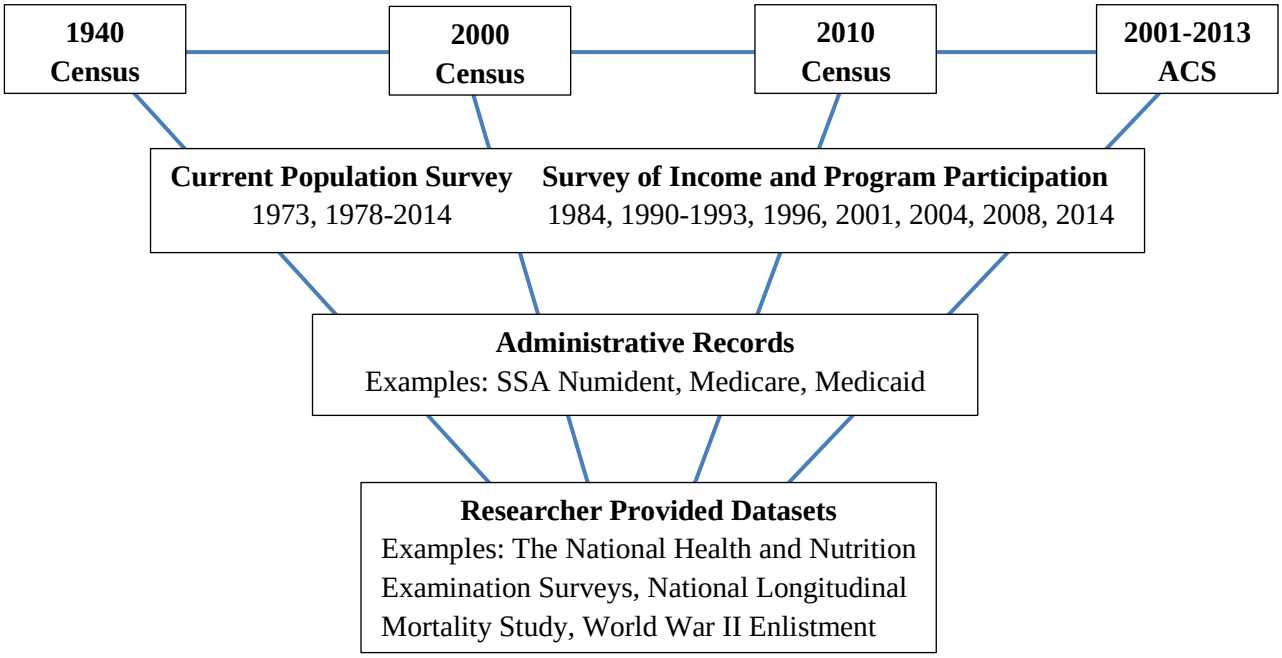


Figure 1 illustrates the build-out potential for the CLIPP infrastructure. We envision that CLIPP researchers could combine any of the datasets displayed in Figure 1. The first tier of Figure 1 represents the core data of CLIPP. These include the 1940 Census, the 2000 Census, the 2010 Census, and the 2001-2013 American Community Surveys. These data enable linkages between millions of individuals observed in multiple decades. To provide additional panel observations, as well as detailed information on income and program participation, surveys such as the CPS and SIPP comprise the second tier of the CLIPP infrastructure. The CPS and SIPP both have explicit longitudinal designs, and CLIPP resources will provide additional data points prior to and following SIPP and CPS surveys.

The third tier of Figure 1 consists of administrative records. The Census Bureau has a long history of using administrative records from federal and state programs. Examples of administrative records that the Census Bureau has been using for decades include the Social Security Administration's Numident file, as well as Medicare and Medicaid databases from the Centers for Medicare and Medicaid Services.

The final tier represents additional data that researchers can provide, depending on the needs of their particular research projects. The CLIPP team will assign PIKs to these files and integrate them into the CLIPP infrastructure.

The sections below describe linkage efforts that have already taken place at the Census Bureau and provide an overview of the record linkage techniques that will be used for the 1940 Census.

2. Large-Scale Efforts to Build Longitudinal Census Data Resources

CLIPP explicitly builds upon several long-standing projects that combine Census data with other files to create longitudinal data resources. In the Census Bureau, these projects include the Longitudinal Employer Household Dynamics (LEHD) program, the SIPP/CPS files linked to Social Security Administration data, and the National Longitudinal Mortality Study (NLMS).

The LEHD integrates federal and state administrative records with census data to create a longitudinal database of employers and employees, with quarterly updates since 1990. The LEHD program uses these data to create statistics on employment, earnings, and job flows, and the microdata have been available for use in Census Bureau Research Data Centers since 2004.

Since the 1990s, the Census Bureau and Social Security Administration have collaborated to link SIPP and CPS records to SSA/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt. Individuals observed in a SIPP or CPS panel can be linked to many years of administrative data, including retrospective and prospective earnings and employer information.

The National Longitudinal Mortality Study (NLMS) provides another example of Census researchers using data linkage to enhance data within the Census Bureau. The NLMS combines CPS ASEC data with death certificate information to determine mortality status and cause of death. Using these data, researchers have produced nearly 100 journal publications.

Outside of the Census Bureau, scholars have linked person records across public (pre-1950) decennial census data (for example, see Ferrie, 1996; Collins and Wanamaker, 2014; Long and Ferrie, 2013; and Abramitzky, Boustan, and Eriksson, 2014). The MPC produced the most comprehensive linkages of individuals across the 1850-1930 censuses in their IPUMS Linked Representative Samples (Ruggles, 2006 and 2011; Goeken, Huynh, Lynch, and Vick, 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.

3. Linkage Techniques at the Census Bureau

Overview of the Person Identification Validation System (PVS)

The Census Bureau uses the Person Identification Validation System (PVS) to assign unique PIKs to person records to facilitate unduplication and record linkage. The PVS uses a probabilistic matching algorithm and processes data through modules, with each module blocking the data in different ways and comparing different fields to find matches (Felligi and Sunter, 1969). Personal identifiers are compared to data in a reference file constructed from the Social Security Administration (SSA) Numident file and other federal-agency administrative data, using different combinations of Social Security Numbers (SSNs), full name, full date of birth, and address (Wagner and Layne, 2014). Records cascade through matching modules, and only those observations that did not receive a PIK pass from one module to the next. All reference file records are available for linkage in each module. Through this process, the PVS appends PIKs to census, survey, and administrative records data.

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 contains 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 includes all transactions associated with each SSN.[1] Consequently, the reference file is large and it is technically infeasible to compare every Numident record to every census record simultaneously. PVS compares census and reference file records within several passes in each module. Each pass alters the comparison of characteristics specified by the data processing staff, using slight variations of the match fields, with subsequent passes permitting 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. PVS employs a string comparator program to measure Jaro-Winkler distances between first and last names in the input and reference files (Winkler, 1995).[2] These distances serve as a metric of how closely two names match, while allowing for some degree of misspelling. For numeric variables, such as year of birth, a maximum acceptable difference between the variable value in the input and reference record is programmable. This also allows for creation of an interval, or band, around year of birth to permit inexact matches. The total score is calculated as the sum of the agreement and disagreement weights attributed to each matching variable (Felligi and Sunter, 1969).

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 the first 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 if 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.

[1] Transactions occur when corrections or name changes are made on a record of a particular SSN. The average number of transactions per SSN is 2.1 (Harris, 2014).

[2] 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.

PVS Modules

PVS employs customized combination of search modules, depending on the characteristics of the input data file. The available PVS modules include the Verification, GeoSearch, Movers, Name, Date of Birth (DOB), Household Composition, and ZIP3 Adjacency modules.

The Verification Module is the first step for all data that contain SSNs. The module matches input records to the reference file by SSN and the compares name and date of birth. If name and date of birth sufficiently agree, 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 a PIK and the household must consist of more than one member. This module considers persons living at the same address as a unit and searches 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, employing parallel processing to identify potential matches. 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. For each household with at least one PIK assigned, a universe of known family members is created from PIKed households 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.

The ZIP3 Adjacency Module expands the geography of the blocks used in the GeoSearch module. This module includes reference records in ZIP3 areas that share a border with the ZIP3 area corresponding to the input record's address. This allows for linkages in cases where there

may be miscodes in the ZIP3 field. The ZIP3 Adjacency module determines matches by the similarity of name, date of birth, sex, and various address fields.

PIK Quality

Research shows that, when a case receives a PIK, quality of the PIK is high. Using Medicare Enrollment Database (MEDB), Indian Health Service (IHS) patient registration files, and commercial data, Layne, Wagner, and Rothhaas (2014) assess the error rate of several modules. For each dataset, they compare a deterministically matched PIK from the administrative record SSN to a PIK from the PVS probabilistic matching to assess the PIK error rate. The MEDB data has a PIK error rate ranging from 0.005 percent for the GeoSearch Module to 1.174 percent for the ZIP3 Adjacency 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 PIKed cases 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 George (2011) find statistically significant differences in PIK rates by race, household size, citizenship, and rural status. Using 2009 ACS data, Mulrow, Mushtaq, Pramanik and Fontes (2011) find substantial geographic differences in PIK assignment. Bond, Brown, Luque, and O'Hara (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 be PIKed, while those in the military and the highly educated are more likely to receive a PIK. The most obvious source of bias in the PIK process is due to the fact that the reference file is built from federal agency records. Anyone who does not have an SSN or 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, Brown, Luque, and O'Hara (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.

4. Using PVS to Link Records with Minimal Personally Identifiable Information

Datasets processed through PVS ideally have a core set of Personally Identifiable Information (PII) including name, date of birth, place of residence, and SSN. Since PVS was designed to link recent Census Bureau surveys to each other and to administrative records, the "Production" version of PVS required significant modifications to work effectively with the limited PII in

historical census data. We developed the Modified Person Identification Validation System (ModPVS) to process the 1940 Census. [3]

Linking Variable in the ModPVS

The ModPVS is designed to PIK 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 Production PVS has well-established techniques for matching on name and date of birth, but new techniques had to be developed to make use of the additional matching variables.

Name is processed using the same methods as the Production PVS system. The text-string comparator built in to PVS compares first and last names. Middle initials and middle names are available for a small number of records and are also compared using PVS's text string comparator.

Date of birth is not available in the 1940 Census; therefore, ModPVS 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 on "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 Production PVS).

Geographic variables used in ModPVS includes birthplace and--for a fraction of records--location in 1940 and 1935. In order to use state or country of birth, Census staff recoded the place of birth information from reference file into the same state/country coding scheme used in the 1940 Census birthplace variable. For SSNs issued within a year of 1940, ModPVS compares the state of residence reported in the 1940 Census to the state of residence that can be 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, ModPVS uses parents' names as linking keys when available on both the 1940 Census and reference file. In the 1940 Census, parents' names can be inferred from relationship information and are linkable to parents' names observed in the Numident. We appended parents names not only for youth, but 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 ModPVS

The ModPVS incorporates the use of additional variables by developing new modules for birthplace (Birthplace 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 a maximum of a two-year age difference between a Census record and a reference file record.

[3] When we refer to the Production PVS, we are describing the official PVS process used to append PIKs to data.

To PIK the 1940 Census, the ModPVS first processes data through the new Birthplace Module, which blocks data by state or country of birth. Next, the ModPVS processes unmatched records through the Name Search Module. This module operates identically to the Name Search Module used in the Production PVS, scoring potential matches using the match fields available in the 1940 Census.

For records that have still not received a PIK, the ModPVS 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 record's SSN was issued and blocks the census by state of residence in 1940.

The Geo1935 Module is identical to the Geo1940 Module except 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 those are still not processed 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.

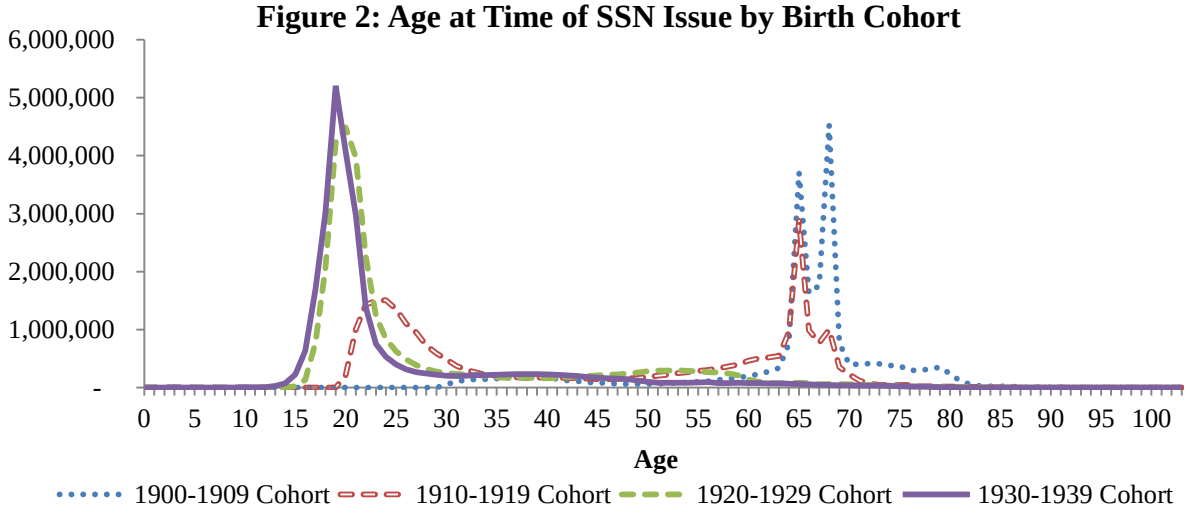
ModPVS Reference File

The effectiveness of the CLIPP historical linkages is reliant on the quality and coverage of the reference file, which is largely comprised of Numident data. Several Numident variables are particularly critical for the ModPVS, including date of birth (used to determine age in April 1940), state where SSN was obtained (for the Geo1940 and Geo1935 Modules), and names (name is a core matching variable, and parents' names used to deduplicate multiple matches and to enhance the reference file with potential maiden names).

Date of birth. The vast majority persons in the Numident born before 1941 have a complete date of birth. Of those born before 1941, approximately 1.8% are missing year of birth, and 2% are missing either month or day of birth. 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.

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 2 charts age at SSN issuance by birth cohort. As Figure 2 shows, those born between 1910 and 1929 often obtained SSNs when entering the labor force. Using the Geo1935 and Geo1940 modules, these states of issuance could be compared to information from the 1940 Census.

Figure 2: Age at Time of SSN Issue by Birth Cohort



Source: 2014 Census Numident

Parents' names. The Numident records parents' first and last names, and ModPVS used these variables for matching. As Table 1 shows, at least one parent's name is recorded in the Numident for 97 percent of those aged 0-9 years in 1940, and for 92 percent 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, we only observe parents' names when the parent and child are co-resident. Due to these coverage issues, the ModPVS's use of parents' names is most successful for those who were under age 20 in 1940.

Table 1: Parent Name Availability in the Numident and the 1940 Census

Age in 1940	Numident				1940 Census			
	Records with at least one parent's name		All births up to 1940		Records with at least one parent's name		All records	
	N	%	N	%	N	%	N	%
0-9	27,420,064	97	28,379,284	100	20,791,049	96	21,762,924	100
10-19	27,897,222	92	30,468,788	100	21,531,326	89	24,234,687	100
20-29	19,662,524	70	28,085,308	100	8,608,777	38	22,791,842	100
30-64	15,241,391	28	54,805,435	100	4,624,241	8	54,768,397	100
65+	1,120,056	47	2,375,517	100	36,823	0	9,073,532	100
Total	91,339,664	63	144,114,332	100	55,592,216	42	132,631,382	100

Source: SSA Numident data and the 1940 Census.

The ModPVS also uses parental names to infer maiden names. Many girls and women in the 1940 Census were enumerated under what would later become their maiden name. Women enumerated under a maiden name in the 1940 Census will only receive an accurate PIK 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, the Minnesota Population Center's Linked Representative Samples focused on men, married couples, and women whose martial status did not change over time. To adjust for this limitation, the ModPVS 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), the ModPVS reference file includes a synthetic "maiden name" record for the woman, with her father's surname in the last name field.

5. Progress on the 1940 Census and Next Steps

The Census Bureau has acquired the digitized 1940 Census records from the Minnesota Population Center. We are currently editing this data for processing by the ModPVS. Census staff have completed most pre-processing for name and birthplace and preliminary PIK assignment is underway.

Once the 1940 data have PIKs, the team will create documentation for the data, and produce reports that assess the representativeness of the PIKed 1940 data, as well as the representativeness of the 1940-2000 linked data, and the representativeness of PIKs in other project files. Along with basic metadata, these linkage reports will become the core of the CLIPP documentation. Technical documentation is also required. For instance, staff will need to document how place of birth is coded, the use of alternate names, birthplace harmonization, and the reliability of SSN area numbers.

The CLIPP team is also conducting tests to estimate the accuracy of ModPVS. To test the proposed techniques to PIK the 1940 Census, we are simulating the 1940 PVS process using Census 2000 long-form data. The 2000 data contain full name, age, place of birth, and PIKs assigned by the Production PVS. Using Census 2000 data allows us to compare the PIKs assigned by the formal Production PVS to those assigned by the ModPVS. Although imperfect, the Production PVS results for the 2000 long-form provide a "truth deck" of sorts, since the Production PVS found these PIKs using detailed PII unavailable in the 1940 Census, including exact date of birth and street address. We simulate the ModPVS on these records by "hiding" this additional detail, and use only linkage variables available in the 1940 Census. By comparing results from the Production PVS and the modified ModPVS, we can calibrate the ModPVS parameters to increase the accuracy of the 1940 process. Testing is currently underway to determine the optimal blocking schemes, scoring, and cutoff values for the 1940 Census.

6. Conclusion

The Core Longitudinal Infrastructure Population Project will create an unparalleled source of uniformly processed linkable census, survey, and administrative data. PIKs will facilitate linkage and unduplication across files, enabling research on population dynamics, migration, life course, social mobility, generational linkages, and socio-economic status. The CLIPP core datasets will include the 2010 Census, 2000 Census, and 1940 Census, as well as survey and administrative records.

Work on CLIPP is underway. The Census Bureau has established a formal internal project to create and document the linked data files, and to establish procedures for external researchers to gain approval to access the files in the Census Research Data Centers. Census staff are now beginning construction of the core record linkages and processing of the 1940 Census.

7. References

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, W. J. & Wanamaker, M. H., 2014. Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data. *American Economic Journal: Applied Economics,* 6(1), pp. 220-52.

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

Fellegi, I. P. & Sunter, A. B., 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, T. A. & Vick, R., 2011. New Methods of Census Record Linking. *Historical Methods,* 44(1), pp. 7-14.

Grusky, D. B., Smeeding, T. M. & Snipp, C. M., 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.

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, C. G., 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, C. G., 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.*

Michelson, M. & Knoblock, C. A., 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.

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., J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek. *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.

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

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