INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
PANEL STUDY OF INCOME DYNAMICS
UNIVERSITY OF MICHIGAN

Technical Series Paper #20-02

Linking 1940 U.S. Census Data to the Panel Study of Income Dynamics: Technical Documentation

July 28, 2020

John Robert Warren[1*]
Fabian T. Pfeffer[2]
Jonas Helgertz[1, 3]
Dafeng Xu[4]

1. Minnesota Population Center, University of Minnesota; 2. Department of Sociology and Institute for Social Research, University of Michigan; 3. Centre for Economic Demography and Department of Economic History, Lund University; 4. Evans School of Public Policy and Governance, University of Washington

* Corresponding email: warre046@umn.edu

This project received support from the U.S. National Institutes of Health (NIH) via grant R01AG050300. It also benefitted from administrative and technical support from the Minnesota Population Center, which receives core infrastructure support from NIH grant 2P2CHD041023, and the Michigan Center on the Demography of Aging, which receives core infrastructure support from NIH grant P30 AG012846.

1. INTRODUCTION

In this document we describe a project to link records from the 1940 U.S. Census to records for individuals in the Panel Study of Income Dynamics (PSID).

The project is part of a larger effort to conduct parallel linkages to the 1940 Census for respondents to the PSID, the Health and Retirement Study (HRS), the Wisconsin Longitudinal Study (WLS), the National Social Life, Health, and Aging Project (NSHAP), and the National Health and Aging Trends Study (NHATS). Each study contains sample members who were alive at the time of the 1940 federal census and were thus enumerated (along with their families and household members). These five ongoing longitudinal studies are central components of America's data infrastructure for interdisciplinary research on aging and the life course; physical and mental health, disability, and well-being; later-life work, economic well-being, and retirement; end-of-life issues, and many other topics. Adding information about sample members from the 1940 Census expands the utility of all five projects and enables important research on the effects of early life social, economic, environmental, contextual, and other factors on subsequent life outcomes. For a longer discussion of the overall linking project, see Warren, Helgertz, and Xu (2020).

Broadly, the project described in this document involved (1) preparing and formatting data files containing cohort members' identifying information; (2) deploying machine learning algorithms to automatically link survey members to their 1940 U.S. Census record; (3) hand linking records that could not be machine linked and hand-verifying a portion of those that could; and (4) documenting the new measures and making them available as part of the PSID's restricted access dissemination systems in a manner consistent with the confidentiality protection of PSID sample members. In this document we describe the linking procedures, explain the structure of the resulting linked files and how they can be accessed, and provide information about linkage rates and the reliability and validity of the links.

This project was led by Dr. John Robert Warren, Dr. Jonas Helgertz, Dr. Dafeng Xu and others at the Minnesota Population Center (MPC) in collaboration with the Fabian Pfeffer and the PSID leadership team at the University of Michigan. It received financial support from the U.S. National Institutes of Health (NIH) via a grant (R01AG050300) to Dr. Warren. It also benefitted from administrative and technical support from MPC and MiCDA, which receive core infrastructure support from NIH (2P2CHD041023 and P30 AG012846, respectively).

Questions regarding these data should be directed to John Robert Warren warre046@umn.edu

2. DATA STRUCTURE AND ACCESS

Who did we try to link?

In principle, we attempted to locate in the 1940 Census PSID sample members who were alive by the time of the decennial enumeration on April 1, 1940. In the PSID, we attempted to link records for 4,101 male sample members who were a PSID reference person ("head") or spouse/partner in any of the PSID waves. For reasons described below, we are still working to link records for PSID women.

What did we produce?

The product of our linking work is a two-variable file that crosswalks (1) HISTID, the permanent person-level identifier in the IPUMS.org version of the complete-count 1940 Census (Ruggles et al. 2020), and (2) the PSID unique identifiers (1968 family interview number and person number; ER30001 and ER30002). This crosswalk is used by PSID staff to link all 1940 Census variables to the PSID sample; due to confidentiality restrictions the crosswalk itself is not accessible by external researchers.

In principle, every variable in the PSID pertaining to people we attempted to link is available to be used in the linkage process. Likewise, every variable in the publicly available complete-count 1940 Census at IPUMS.org is available to be linked to PSID survey data. This includes individual level census data (for PSID sample members but also anyone they were living with in 1940), family and household level data, and geographic/administrative information (e.g., state, county, enumeration district, street address). However, as described below, and to protect PSID sample members' identities and privacy, some information is suppressed in files available to researchers.

How to access survey data linked to the 1940 Census

To protect sample members' identities and privacy, researchers need to access linked census and survey records via the PSID's restricted data access and licensing protocols. This means that linked files will be available remotely via a secure enclave server at the University of Michigan after appropriate data use agreements and security protocols have been established. To apply to access linked 1940-PSID data, please read the instructions here.

Structure of the Linked 1940-PSID Files

The linked 1940-PSID data are contained in four different SAS data files; the file to which a researcher would have access will depend on their analytic needs.

1. PSID_1940_HHFile

This file contains the household level information from the 1940 Census for those 1940 Census households who have been linked to PSID Persons. It contains 1,706 records and 50 variables. Variables include information on household size and structure and socio-economic circumstances (e.g., ownership of dwelling, house value).

In addition to the Census variables, we have added a release number and two fields for ID68 (ID68MEN1 and ID68MEN2). These fields correspond to the PSID key variable 1968 INTERVIEW NUMBER (ER30001). The second mention of ID68 was added because a small number of PSID Persons who do not share the same 1968 Interview Number do appear in the same 1940 Census household. There are 5 pairs of individuals in this situation and another 6 pairs of individuals who do share both the Census household as well as the same 1968 Interview Number.

We have also added the variable FAM_HHID, containing a random household ID starting with 1000001. The FAM_HHID number links the HHFile with the PSID Person file as well as the Non-PSID Person File. FAM_HHID replaces the Census variables Serial and SerialP to protect the anonymity of our respondents. There is a one-to-many match between the FAM_HHID and each of the Person Level files as multiple persons may reside in the same Census household.

2. PSID_1940_PSIDPersonFile

This file contains the individual level information from the 1940 census for those PSID Persons who have been linked to a 1940 Census household. It contains 1,717 records and 116 variables. Variables include information on relationship to the household head, age, sex, race and ethnicity, marital status, birth place, citizenship status, school enrollment and educational attainment, place of birth, place of residence in 1935, labor force status, industry, occupation, working hours, and income.

In addition to the Census variables, we have added a 1968 Family Identifier labeled ID68 which corresponds to the Individual File variable (ER30001) and Person Number labeled PN which corresponds to ER30002. This file can be merged to the PSID_1940HH file using the variable FAM_HHID.

3. PSID_1940_NonPSIDPersonFile

This file contains all persons who live in a 1940 Census household that has been linked to a PSID person but who have not been identified as PSID individuals themselves. It contains 7,555 records and the same variables as included on the PSIDPersonFile with the exception of the two PSID specific person identifiers.

4. PSID_1940_HHGeoFile

An additional file containing geographic identifiers is available based on separate permission within the restricted use data contract. This file contains the 1940 Census enumeration district variable. Because of the level of precision of enumeration districts, researcher must provide an explicit and detailed justification for exactly how and why the research will benefit from having access to these data. Your Research Plan must include a description of the data that you will merge to the enumeration district.

3. LINKING PROCEDURES

The objective of the project—to identify PSID respondents in the 1940 full count census—was achieved using a combination of a machine learning algorithm and hand linking protocols. In this section, we present an overview of our linking procedures.

Data Cleaning

Machine linkage algorithms compare text strings and quantify their similarity; in our case, strings from the 1940 Census were compared to those from the PSID. However, formatting

Issues - whether strings are in upper case or lower case, whether spaces have been removed, whether punctuation is included, how abbreviations are treated, etc.—influence the results of machine record linkage. The strings “North Dakota” “N. Dakota” and “northdakota” may cognitively seem the same to a human being, but they would be treated as very different strings by a computer. So, we began by standardizing place names and given names (e.g., “Willie” and “Wm.” Were transformed into “William”).

Machine Linking Algorithm

Our machine linkage algorithms considered PSID sample members’ first names, last names, and years of birth (inferred from their ages) in attempting to locate corresponding individuals in the 1940 Census. As described below, other information about PSID sample members—which is available for only portions of the sample—was used in the process of hand linking and hand verification. Because the PSID records of the target women included here do not include their last names at birth, and because most women in these birth cohorts married and changed their last names upon marriage, we did not attempt to link PSID women. (However, we are currently exploring the possibilities to augment our data with our administrative data to attempt this linkage for women).

To match PSID sample members’ records to the 1940 Census we first defined the universe of potential matches. To make the task computationally tractable, we restricted the population of potential matches to PSID sample members in the 1940 Census to records that displayed identical or similar characteristics on features that should be consistent over time---in this case, year of birth. For example, when attempting to find the 1940 Census record for Michael Corcoran—born in 1917 according to PSID records—we limited the population of potential matches in the 1940 Census to males born between 1914 and 1920. Because age was reported in the 1940 Censuses rather than year of birth, and because of reporting inaccuracies, we allowed for deviations of ±3 years in birth year across data sources.

To identify the correct record from among all possible records in the 1940 Census—e.g., the correct “Michael Corcoran” in 1940 from among all the possible Michael Corcorans—we employed a machine learning algorithm that performed probabilistic record linking techniques. Briefly, we trained a computer algorithm to recognize patterns in a dataset of potential matches that were consistent with a true match. We used a modification of Feigenbaum’s (2016) probit regression approach, which—like other methods of supervised machine learning—requires input from training data. The training data represent a subsample of the population that one wishes to link, but where links have been declared by a trained human to ascertain that confirmed links are as accurate as possible. We used these training data to calibrate the linking algorithm and to evaluate how well it performed in declaring matches and avoiding false positives matches.

To construct the training data, we randomly selected 500 age-appropriate men from the PSID who we then attempted to link by hand to the relevant universe of possible 1940 matches. Here, the universe of potential matches was limited to cases where name similarity scores (Jaro 1989; Jaro 1995; Winkler 1990; Winkler 2006) were at least 0.8, in addition to being of the same gender and with year of birth within +3/-3 years. For PSID individuals for whom state of birth is available, we also restrict the universe of potential matches to 1940

Census individuals born in that same state. As a result of the aforementioned blocking criteria, the number of PSID individuals in the training data was reduced to 483. We assessed all potential matches by hand---and using all available identifying information in the PSID. Using these procedures, and carefully guarding against false positive matches, we were able to manually declare 23.8 percent of the training data sample as uniquely matched.

To calibrate our linking algorithm, we implemented a “train-test-split” procedure using our training data (in which true matches are known). In the first part of the procedure, we split our training data into two equally sized parts. To train the algorithm, we fit a probit regression model on one-half of the sample, and then evaluated its out-of-sample performance on the other. Results from the model informed the algorithm as to which, if any, of the universe of possible matches should be considered a valid link. The algorithm declared a unique link based on (1) the greatest similarity between any 1-to-1 match (technically the predicted probability based on the probit regression estimates) and (2) the relative difference between the best and second-best possible match. By looping multiple times over a range of realistic values on both parameters, we were able to choose values for (1) and (2) that optimized the overall performance of the linking algorithm. We judged overall performance by the algorithm’s ability to minimize false positives (incorrectly linked cases) while maximizing true positives (correctly linked cases) and true negatives (correctly unlinked cases).

In selecting thresholds for declaring matches in our data, we used the Matthew’s Correlation Coefficient (MCC) which is an especially useful measurement for two-class data where the classes are not well balanced (Chicco 2017). This is certainly the case in our application in which the 483 individuals in the training data were typically linked several potential matches in the 1940 Census. The MCC, in Equation 1 below, compares the predictions of the algorithm to all possible outcomes (true/false positives/negatives) and provides a single metric (ranging from -1 to +1) to be used to select which thresholds to use. The formula is as follows, where *TP* represents true positive, *TN* repre. sents true negative, *FP* represents false positive, and *FN* represents false negative:

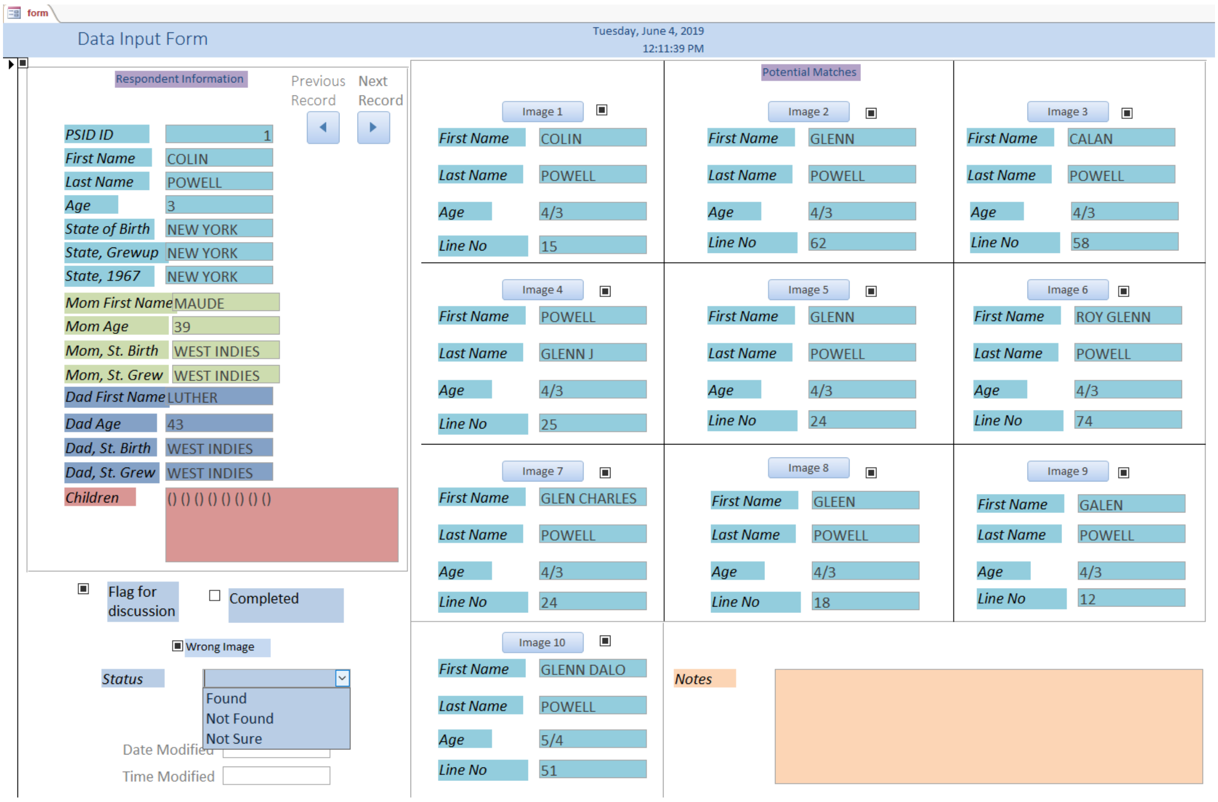
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{1}$$

After training the machine learning algorithm as described above, we applied it to the full PSID sample of 4,101 age-appropriate men. For each person, this yielded a set of possible matches; as noted above, a unique match was declared when there was one and only one high quality match in the 1940 Census file.

Hand Linking and Verification Procedures

To verify the work of machine linking algorithm and to attempt to confidently determine high quality matches when the machine algorithm failed to do so, we next performed a set of hand linkages.

For most of the 4,101 cases, trained hand linkers were presented with a screen like the one below. On the left side is information about the focal PSID sample member (who in this fictional example is General Colin Powell). The hand linker is provided information—when available—about the PSID person's first name, last name, age (in 1940), state of residence while growing up, state of residence upon entering the PSID, mother's and father's name, mother's and father's age (in 1940), mother's and father's states of birth, and mother's and father's states of residence upon the focal person entering the PSID. Note that most of this information was unavailable in most cases (which is why it was not used in the machine linking algorithm).



Data Input Form

Tuesday, June 4, 2019
12:11:39 PM

Respondent Information

PSID ID	1
First Name	COLIN
Last Name	POWELL
Age	3
State of Birth	NEW YORK
State, Grewup	NEW YORK
State, 1967	NEW YORK
Mom First Name	MAUDE
Mom Age	39
Mom, St. Birth	WEST INDIES
Mom, St. Grew	WEST INDIES
Dad First Name	LUTHER
Dad Age	43
Dad, St. Birth	WEST INDIES
Dad, St. Grew	WEST INDIES
Children	() () () () () () () ()

▣ Flag for discussion   □ Completed

▣ Wrong Image

Status: Found / Not Found / Not Sure

Date Modified

Time Modified

Potential Matches

	Image 1 ▣	Image 2 ▣	Image 3 ▣	Image 4 ▣	Image 5 ▣	Image 6 ▣	Image 7 ▣	Image 8 ▣	Image 9 ▣	Image 10 ▣
First Name	COLIN	GLENN	CALAN	POWELL	GLENN	ROY GLENN	GLEN CHARLES	GLEEN	GALEN	GLENN DALO
Last Name	POWELL	POWELL	POWELL	GLENN J	POWELL	POWELL	POWELL	POWELL	POWELL	POWELL
Age	4/3	4/3	4/3	4/3	4/3	4/3	4/3	4/3	4/3	5/4
Line No	15	62	58	25	24	74	24	18	12	51

Notes

On the right side of the figure are up to 10 possible matches in the 1940 Census. These were the (up to 10) best matches generated by the machine linking algorithm above; hand linkers were not told which (if any) of the possible matches were declared by the machine to be a valid match and the order of the possible matches was randomized. Note that hand linkers were provided with each possible match's name, age (in 1940), and a line number. The line number corresponds to the line on the 1940 Census enumeration form on which the person appears; the hand linker was then able to click on the "Image" button to view the actual 1940 Census enumeration form. Hand linkers proceeded by comparing the information about possible matches provided and on the enumeration form with information about the focal PSID sample member. Being able to see the actual enumeration forms allowed hand linkers to consider additional information like place of birth, place of residence in 1940, parents' ages, and parents' places of birth, but also possible transcription errors arising, e.g., from difficult-to-read handwriting of a Census enumerator. If hand linkers determined that one—and only one—of the 10 presented options was a valid and correct match, they checked "found" and "completed" in the lower left of the screen. If they could not decide between two possible matches with the information provided, they declared "not found." In cases in which hand linkers were unable to reach a decision, they marked "not sure" and the case was reviewed by one of the authors.

Not all the 4,101 PSID cases were hand linked. For cases in which the machine linking algorithm declared one and only one very high-quality match, there was no need for hand verification—nothing the hand linker could see would reduce confidence in the quality of the match. For cases in which the machine linking algorithm declared zero reasonable-quality matches, there was also no need for hand verification—none of the cases could plausibly be a valid match. Thus, the hand linkers mainly focused on cases in which there were two or more reasonably plausible matches. Their job was to adjudicate, if possible, between them.

To assess the reliability and validity of the overall linking efforts, we further implemented two procedures in the hand linking phase. First, we assigned hand linkers to attempt to hand link a random subset of the cases that the machine linkage algorithm declared had one and only one high-quality match. This allowed us to assess validity—how often the hand linker selected (by hand) the same match in the 1940 Census as the machine selected (via computer algorithm). Second, we assigned a randomly selected subset of hand linking cases to multiple hand linkers. This allowed us to assess the reliability of the hand linkers' decisions.

All hand linkers met regularly with the authors to ask questions; to review problematic or instructive cases; to review reliability and validity statistics; and to discuss progress. When necessary, hand linkers whose work was insufficiently valid or reliable were retrained or replaced.

4. LINKING RESULTS

Linkage Rates

Of the 4,101 age appropriate PSID men we attempted to link, we were able to confidently declare high-quality matches for 1,717 (or 41.9%) of them to the 1940 Census.

In most cases, linking occurred only based on PSID sample members' names and years of birth. When other information was available, linkage rates were higher (because it became possible to adjudicate between multiple possible matches). For example, when information about state of birth was available, 54.7% of cases were matched. When one or more parents' names were available, 50.9% of cases were matched.

Reliability and Validity

As noted above, we assessed the validity of the matches by assigning to the hand linkers a random subset of cases for which the machine algorithm declared there to be one and only one high-quality match. For these cases, the machine algorithm and the hand linker arrived at the same decision about a match 75% of the time. However, in most cases disagreements between the machine algorithm and the hand linker were due to the hand linker's reluctance to declare any match from among the options available to them. When hand linkers did choose a match from among the options available to them, they chose the same match as the computer algorithm 94.9% of the time.

To assess the reliability of hand linkers' efforts, we assigned a random subset of hand linking cases to two hand linkers who then independently worked the cases. For these cases, the two independent hand linkers reached the same conclusion 72.7% of the time. However, most of the disagreement in these instances was due to one linker declaring a match and the other declaring no match. When both linkers declared matches, they chose the same match 98.3% of the time.

In short, the validity and reliability of the linking efforts are high. Recall also that the machine linking algorithm was trained to minimize false negatives (while also maximizing overall linkage rates).

5. REFERENCES

Chicco, D. 2017. "Ten quick tips for machine learning in computational biology." *Biodata Mining* 10.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking." Cambridge, MA: Department of Economics, Harvard University.

Jaro, M. A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84(406):414-20.

—. 1995. "Probabilistic Linkage of Large Public-Health Data Files." *Statistics in Medicine* 14(5-7):491-98.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. " IPUMS USA: Version 10.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.

Warren, John Robert, Jonas Helgertz, and Dafeng Xu. 2020. "Linking 1940 U.S. Census Data to Five Modern Surveys of Health and Aging: Technical Documentation." Available as MPC Working Paper #2020-01 at https://pop.umn.edu/research/working-papers. Minneapolis, MN: Minnesota Population Center, University of Minnesota.

Winkler, W.E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." American Statistical Association, 1990 Proceedings of the Section of Survey Research Methods, 354-359. http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf.

Winkler, William E. 2006. "Overview of Record Linkage and Current Research Directions." Research Report Series (Statistics #2006-2), available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf. Washington, D.C.: U.S. Census Bureau.