CenSoc WWII Army Enlistment Dataset*

Anna Wikle† Maria Osborne ‡

August 22, 2023

Abstract

In this technical report, we introduce the CenSoc WWII Army Enlistment Dataset. This harmonized file was constructed from the most informative parts of World War II era army enlistment records that were digitized and published by the National Archives and Records Administration and the U.S. Census Bureau. The CenSoc WWII Army Enlistment Dataset contains information on 24 variables for over 9 million records of enlistees circa 1938-1947. This report details the composition of these records, and describes steps taken to clean the raw data file to produce this harmonized data set. We also describe datasets constructed by linking the CenSoc WWII Army Enlistment Dataset to U.S. mortality records and census records.

*For helpful discussions and feedback, we thank members of the CenSoc working group. Research reported in this technical report was supported by the National Institute of Aging grant R01AG05894 and R01AG076830.

†Department of Epidemiology/Biostatistics, University of California, Berkeley. annawikle@berkeley.edu.

‡Department of Demography, University of California, Berkeley. mariaosborne@berkeley.edu.

Contents

1 Introduction

American involvement in World War II constituted the largest war effort in the nation's history, with nearly 9% of the resident population serving in active duty military roles at peak enrollment (Clever and Segal, 2012). Beginning in the 1990s, the U.S. Census Bureau and the National Archives and Records Administration (NARA) digitized microfilmed Army enlistment records from the World War II era, eventually creating a single electronic file with individual-level data on over 9 million United States Army enlistees circa 1938-1946.

NARA makes these data freely available to the public, and individual records are searchable through their Access to Archival Database. However, the presence of large amounts of scanning errors and non-standard coding schemes present difficulties for researchers seeking to analyze the full data file. The CenSoc WWII Army Enlistment Dataset cleans and harmonizes raw data from NARA in order to create a dataset that is readily workable and linkable to other data sources. Variables from enlistment records include information on sociodemographic data such as education level and race, in addition to information related to military position and rank. The inclusion of body height and weight make this dataset a uniquely large source of anthropometric data. The availability of names, birth years, and birthplaces allow us to link the data to Social Security Administration mortality records and to the complete-count 1940 Census. This report details how the original NARA data file was cleaned and harmonized to produce the CenSoc WWII Army Enlistment Dataset. This file contains over 9 million records and 24 variables (Table 1).

2 History and Contents of WWII enlistment records

2.1 History and public release of NARA data

World War II era military enlistment records originally existed as IBM punch cards. In 1947, the Personnel Services Support Division of the Adjutant General's Office microfilmed these punch cards, creating the "Microfilm Copy of the Army Serial Number File, 1938–1946," and then destroyed the punch cards (Hull, 2006). NARA acquired the rolls of microfilm in 1959, and in 1992 contracted with the Census Bureau for a special conversion project which later

Variable	Label
serial_number	Army Serial Number
byear	Year of Birth
fname	First Name
mname	Middle Name
lname	Last Name
sex	Sex
date_of_enlistment	Date of Enlistment
bpl	Place of Birth
residence_state	State of Residence at Enlistment
residence_county	County of Residence at Enlistment
place_of_enlistment	Place of Enlistment
education	Education
grade_code	Army Grade (Rank)
branch_code	Army Branch
term_of_enlistment	Term of Enlistment
race	Race
citizenship	Citizenship
civilian_occupation	Civilian Occupation
marital_status	Marital Status
height	Height at Enlistment (inches)
weight_before_march_1943	Weight at Enlistment (pounds)
weight_or_AGCT	Weight (pounds) or AGCT score
component	Army Component
source	Source of Army Personnel

Table 1: Variables in the CenSoc WWII Army Enlistment Dataset

resulted in the creation of the series "Electronic Army Serial Number Raw Files, 1994-2002."

The Census Bureau converted microfilmed punch card images to a computer-readable format, noting locations where punch cards were unreadable, and scanning some cards up to 10 times if any data could not be extracted upon a single read. In order to prepare the data for entry into NARA's Access to Archival Databases (AAD) resource, which allows users to search for individual records using various search fields, NARA created the "Electronic Army Serial Number Merged File, 2002", which contains a single electronic record for each punch card. This involved "merging" multiple readings of the same card by collapsing the first two scans of each card into a single "best guess" record, ignoring any subsequent scans. A more detailed history of these data conversion processes is available from Hull (2006).

To create the CenSoc WWII Army Enlistment Dataset, we use the Electronic Army Serial Number Merged File, rather than returning to the Census Bureau's raw data files to create our own "best guess" from multiple readings of the same records. In this report, references to NARA raw data or NARA merged data refer to the Electronic Army Serial Number Merged File, 2002.

2.2 Structure and content of original enlistment records

The NARA data comprise approximately 9 million records of men and women who enlisted in the United States Army, including the Women's Army Auxiliary Corps and the Army Air Corps (*Electronic Army Serial Number Merged File, ca. 1938 - 1946*, 2002). The variables collected include serial number, name, state and county of residence, place of enlistment, date of enlistment, army grade (rank), army branch, term of enlistment, place of birth, year of birth, race, citizenship, education, civilian occupation, marital status, height and weight (before approximately March 1943), component (e.g, national guard, army reserves), and source (e.g. civil life or other army faction).

Overall, the records represent the majority of service members who enlisted during this time period. However, there are gaps in records, and about 15% of microfilm images within valid serial number ranges were unreadable. The known gaps in serial numbers are detailed in NARA's series description of the Electronic Army Serial Number Merged File. The merged NARA data file has 9,200,232 records, but 160,392 of these records simply mark a location

in the microfilm where one or more punch cards were not readable. We have removed these, leaving a sample of 9,039,840 enlistees.

3 CenSoc WWII Army Enlistment Dataset

3.1 Cleaning and harmonizing the NARA enlistment records

NARA merged enlistment records are rife with invalid data and contain a number of non-standard codes and coding schemes. Where possible, we have harmonized variable coding schemes to align with those used by IPUMS-USA (Ruggles et al., 2020), such as those variables for race and educational attainment. For variables which had no IPUMS analog, we generally replace non-standardized coding schemes with simple numeric codes. Some categorical variables are left relatively untouched, as they contain thousands of potential values. Valid codes from original NARA data are collected in technical documentation for the Electronic Army Serial Number Merged File. A detailed description of the treatment of individual variables is as follows:

- **Names** Raw name strings were separated to create three variables for first, middle, and last names (fname, mname, and lname). Names are cleaned to remove titles (e.g., "Dr.") and non-alphabetic characters, and are converted to lowercase. Last names are standardized by removing whitespace (e.g., the surnames "St. John", "StJohn", and "St John", are all coded as "stjohn").
- **Dates** Year of birth and date of enlistment were processed and filtered to include valid dates. We retain birth years from 1860-1930 and enlistment dates from 1938-1947 Figure 1 shows the number of enlistments by month in the harmonized enlistment data.
- **Site of enlistment** Site of enlistment was left as-is, as this variable contains over 3000 location codes with no analogous coding existing in IPUMS data or elsewhere.
- **State and county of residence** State of residence was recoded to align with IPUMS-USA detailed birthplace codes. Original state of residence information used multiple codes for some states of birth, which were condensed into a single code for each U.S.

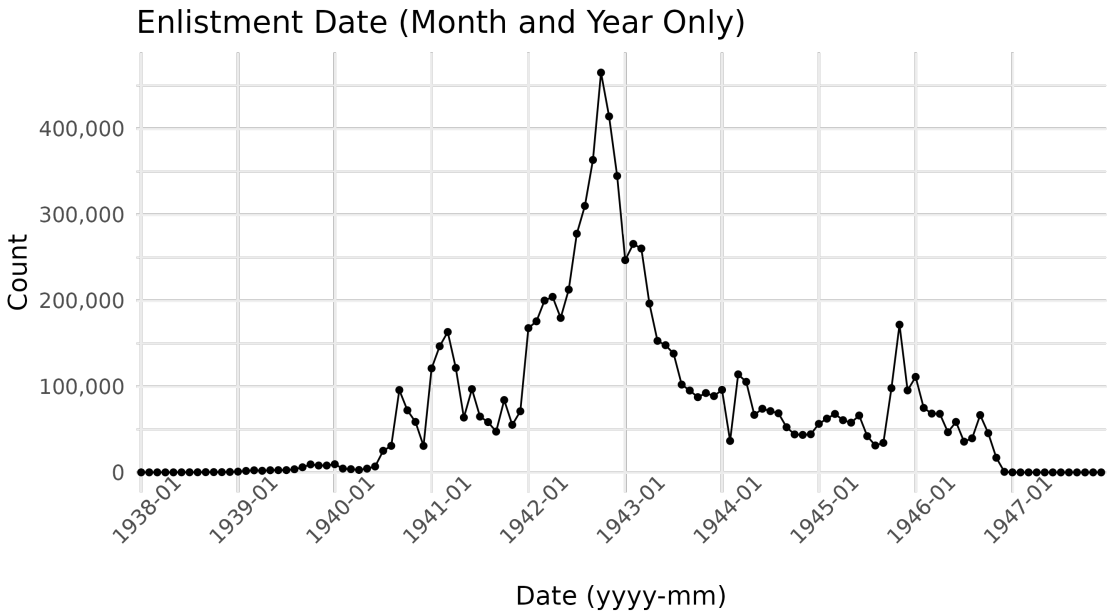

Figure 1: Counts of enlistment numbers over time

state. Places outside the United States, Alaska, Hawaii, and Puerto Rico were removed. Some original residence codes referred both to a U.S. state and a foreign nation: for example, the code **N1** corresponds to "INDIANA, LIMITED SERVICE or (DANZIG or GERMANY)." We select the U.S. state as the place of residence in these cases, as the vast majority of individuals with ambiguous residence state codes have valid domestic residence county codes. County of residence is left as-is, but removed for enlistees residing outside the United States (including Alaska and Hawaii, which were not yet states). County codes published by NARA are unique only within states and generally equivalent to FIPS county codes. However, they may not align exactly with modern FIPS codes due to county splits, merges, name changes, etc. occurring since the 1940s. Invalid county codes may remain in the data.

- **Birthplace** Raw nativity (birthplace) codes from NARA were converted to IPUMS-USA detailed birthplace codes from 1940 to standardize locations and facilitate matching. However, there were several instances where a NARA place of birth codes corresponded to multiple locations, without a direct IPUMS analog. Where a single code corresponded to multiple countries, the most populous location was used to select an appropriate IPUMS code. For example, NARA code **K6** corresponds to "AUS-

TRALIA or BISMARK ISLANDS or BRITISH AUSTRALASIA AND OCEANIA or FIJI ISLANDS or NEW GUINEA or NEW ZEALAND or OTHER PACIFIC BRITISH ISLANDS or PAPUA or SOLOMON ISLANDS or TASMANIA", and was recorded to the IPUMS code **7000** meaning "Australia and New Zealand".

- **Military-specific variables** Grade (rank) codes, branch codes, source, term of enlistment, and component were cleaned to remove invalid codes and characters. Some codes that originally indicated equivalent values were condensed into single codes. In NARA data, army grade and branch consisted of two separate fields: an alphabetic character code (such as "PVT" to indicate the rank of private) and a numeric code. As the character codes are highly prone to misspellings, we have removed these and only publish the numeric codes.
- **Race and citizenship**. Race and citizenship were initially combined into one field, which we have separated into two distinct variables. The race variable was harmonized with IPUMS coding.
- **Education and marital status** Education and marital status were harmonized with IPUMS coding. Raw marital status data included information on whether enlistees had dependents or not, but when harmonizing with IPUMs the distinction between persons with and without dependents was lost.
- **Height** Raw height and weight data present particular problems, as other information may have been recorded in the same location on enlistment punch cards, such as army occupation. Height was filtered based on a reasonable range of numbers permissible for enlistment in the military, 60 - 78 inches (Karpinos, 1958). After February 1943, the distribution of raw data in this field changes drastically, and likely records information other than height, so we have removed data after this date. Height is removed for enlistees with unknown date of enlistment.
- **Weight and AGCT score** Weight and Army General Classification Test (AGCT) scores were recorded in the same location on enlistment punch cards. Prior to March 1943, this variable likely corresponds to weight, and after this time frame it logs either weight or AGCT score based on the place of enlistment. There is a period of months in early 1943 when different sites of enlistment may have been recording weight or AGCT

score and the data are ambiguous (Ferrie, Rolf and Troesken, 2012). Because of this, we have divided raw weight data into two separate variables. One contains weights for enlistees who enlisted prior to March 1943. We have trimmed weights to a minimum of 105 pounds (the lightest acceptable weight for the shortest acceptable inductees), and a maximum of 500. For persons enlisted in March 1943 or later, we create a variable that may contain weight or AGCT score. We remove zeros and values over 500 to reduce the amount of data representing information other than weight or AGCT score, but this field is otherwise left unsorted and uncleaned. Like height, weight/AGCT information is removed for enlistees with unknown date of enlistment.

- **Civilian occupation** Civilian occupation was cleaned to remove any information that did not correspond to a valid occupation code.
- **Sex** While the sex variable is not directly in original NARA data, we use army branch information to infer it. Specifically, if an individual's numeric branch code or alphabetic branch code corresponds to the Women's Army Corps, we declare that person to be female. All other persons are assigned male. Using this approach, we identify approximately 133 thousand women in the enlistment records. However, there is potential for error in this designation due to missing branch information or disagreements between army branch numeric codes, branch alphabetic codes (which we do not publish), and army component information. Thus, sex may not always correspond exactly to army branch and component information in the published data set.

We note that some original variables from the merged NARA data have been removed, including alphabetic branch and grade codes; fields which contained no substantive information; and fields containing information on the digitization process of records, such as box number of microfilmed punch cards.

4 Linkages to other datasets

These enlistment records may be nominally linked to other data using additional information such as birthplace and birth year. We match male enlistees to the Berkeley Unified

Numident Mortality Database (BUNMD) (Goldstein et al., 2023) and the Social Security Death Master File (DMF), creating the CenSoc Enlistment-Numident (N=1,692,279) and CenSoc Enlistment-DMF (N= 1,854,783) mortality files, respectively. The records are linked using the ABE matching algorithm used to create other CenSoc datasets, as described in Abramitzky et al. (2021, pp. 871-872). We match on standardized first name, last name, birth year, and sex. Standardizing first names involves fixing common misspellings and replacing nicknames with "standard" names, such as "Bobby" with "Robert". Records are matched on birthplace between the BUNMD and enlistment records, but this variable is not available in the DMF. In order to minimize false matches, we use a conservative variant of the ABE linking process that requires matches to be unique within a five-year window on year of birth ($\pm$ 2 years).

We have also linked the CenSoc WWII Army Enlistment Dataset directly to complete-count 1940 Census data from IPUMS (Ruggles et al., 2020), establishing 2,573,678 matches between enlisted men and the 1940 Census. Enlistment records that are linkable to the census are indicated in the CenSoc Enlistment-Numident and CenSoc Enlistment-DMF datasets. Researchers can attach the 1940 Census variables to these datasets using the *histid* variable. Variables in the CenSoc Enlistment-Numident file are listed in Table 2, variables for the CenSoc Enlistment-DMF file are listed in Table 3, and variables for linked 1940 Census-Enlistment file are listed in Table 4.

All linked enlistment data sets contain only men, as smaller numbers of women are present in enlistment records and linkage rates are low (under 20% of male enlistment records are linkable to the BUNMD). As marital status is recorded in enlistment records, however, it is theoretically possible to link women to data where information on birth/married names is available.

Variable	Label
id	Unique identifier
sex	Sex
bpl	Place of Birth
byear_numident	Year of Birth (Numident)
bmonth_numident	Month of Birth
dyear_numident	Year of Death
dmonth_numident	Month of Death
death_age_numident	Age at Death (years)
race_first_numident	Race on First Application
race_first_cyear_numident	First Race: Application Year
race_first_cmonth_numident	First Race: Application Month
race_last_numident	Race on Last Application
race_last_cyear_numident	Last Race: Application Year
race_last_cmonth_numident	Last Race: Application Month
zip_residence_numident	ZIP Code of Residence at Time of Death
socstate_numident	State where Social Security Number was Issued
age_first_application_numident	Age at first Social Security application
byear_enlistment	Year of Birth (Enlistment)
date_of_enlistment_enlistment	Date of Enlistment
residence_state_enlistment	State of Residence at Enlistment
residence_county_enlistment	County of Residence at Enlistment
place_of_enlistment_enlistment	Place of Enlistment
education_enlistment	Education
grade_code_enlistment	Army Grade (Rank)
branch_code_enlistment	Army Branch
term_of_enlistment_enlistment	Term of Enlistment
race_enlistment	Race (Enlistment)
citizenship_enlistment	Citizenship
civilian_occupation_enlistment	Civilian Occupation
marital_status_enlistment	Marital Status
height_enlistment	Height at Enlistment (Inches)
weight_before_march_1943_enlistment	Weight at Enlistment (Pounds)
weight_or_AGCT_enlistment	Weight (Pounds) or AGCT score
component_enlistment_enlistment	Army Component
source_enlistment	Source of Army Personnel
HISTID	1940 Census Historical Unique Identifier

Table 2: Variables in linked Enlistment-Numident data

Variable	Label
id	Unique Identifier
sex	Sex
byear_DMF	Year of birth (Numident)
bmonth_DMF	Month of birth
dyear_DMF	Year of death
dmonth_DMF	Month of death
death_age_DMF	Age at death (Years)
byear_enlistment	Year of Birth (Enlistment)
date_of_enlistment_enlistment	Date of Enlistment
bpl_enlistment	Place of Birth
residence_state_enlistment	State of Residence at Enlistment
residence_county_enlistment	County of Residence at Enlistment
place_of_enlistment_enlistment	Place of Enlistment
education_enlistment	Education
grade_code_enlistment	Army Grade (Rank)
branch_code_enlistment	Army Branch
term_of_enlistment_enlistment	Term of Enlistment
race_enlistment	Race
citizenship_enlistment	Citizenship
civilian_occupation_enlistment	Civilian Occupation
marital_status_enlistment	Marital Status
height_enlistment	Height at Enlistment (Inches)
weight_before_march_1943_enlistment	Weight at Enlistment (Pounds)
weight_or_AGCT_enlistment	Weight (Pounds) or AGCT score
component_enlistment_enlistment	Army Component
source_enlistment	Source of Army Personnel
HISTID	1940 Census Historical Unique Identifier

Table 3: Variables in linked Enlistment-DMF data

Variable	Label
HISTID	1940 Census Historical Unique Identifier
byear	Year of Birth
sex	Sex
date_of_enlistment	Date of Enlistment
bpl	Place of Birth
residence_state	State of Residence at Enlistment
residence_county	County of Residence at Enlistment
place_of_enlistment	Place of Enlistment
education	Education
grade_code	Army Grade (Rank)
branch_code	Army Branch
term_of_enlistment	Term of Enlistment
race	Race
citizenship	Citizenship
civilian_occupation	Civilian Occupation
marital_status	Marital Status
height	Height at Enlistment (Inches)
weight_before_march_1943	Weight at Enlistment (Pounds)
weight_or_AGCT	Weight (Pounds) or AGCT score
component	Army Component
source	Source of Army Personnel

Table 4: Variables in linked 1940 Census-Enlistment data. Note: all variables included in this dataset, except for HISTID, are taken from enlistment records. Variables such as marital status and residence state reflect information at time of enlistment, not the 1940 Census. No census variables are included in this dataset and must be attached by merging the data with IPUMS 1940 Census data on HISTID.

5 Considerations and Recommendations for researchers

5.1 Limitations

There are few key considerations for working with the CenSoc Army Enlistment Dataset. The original records have a number of data quality issues, such as those arising from inconsistencies in data collection and scanning errors during the digitization process. Nonsensical values for each variable—such as implausibly high values for height or unintelligible strings in name variables—are common. We have removed a great deal of obviously incorrect information present in NARA data, but researchers should be aware that some errors still persist.

The weight and the Army General Classification Test (AGCT) variables are particularly problematic. We have left the weight and AGCT variable unsorted for individuals who enlisted after February 1943. After this point, whether the data contained in the field indicates AGCT score or body weight likely varies systematically by site. Researchers may consult Ferrie, Rolf and Troesken (2012) for possible strategies for detangling these two variables. Further, enlistment records may record information other than weight or AGCT score in this field, such as army occupation.

Beyond problems with individual variables, approximately 1–2 million enlistment records are excluded from the dataset, as a significant proportion of microfilmed punch cards were unreadable and could not be converted to a computer-readable format. NARA documentation does not suggest that these scanning problems systemically exclude certain types of records, but it is not known if readable records comprise a representative set of all enlistees.

Finally, army inductees are of course not perfectly representative of the general American population. Men who enlisted in the armed forces were subject to physical standards, cognitive requirements, racial discrimination, and other forms of selection. Because of these issues, we make no assumption about the population represented by available enlistment records, and do not construct statistical weights for these datasets. Because of the low number of women enlistees, and low linkage rates between enlistment records and other datasets, we only publish linked enlistment data with men.

5.2 Choosing a dataset

CenSoc publishes four datasets with Army Enlistment data: one standalone file and three that contains links to other data. Here, we provide some guidance for researchers choosing which dataset to use for their own research.

The standalone CenSoc WWII Army Enlistment Dataset contains a wealth of information on army enlistees, including birthplace, education, marital status, height, and weight. It is the largest of the enlistment datasets and includes both men and women, but may not contain all information of interest to researchers. Each linked dataset, while much smaller due to unlinkable records, contains additional information and has various advantages depending on desired use. For researchers who simply want to attach additional 1940 Census data to enlistment records, such as information on enlistee income or household structure, we recommend using the CenSoc Enlistment-Census-1940 file.

Researchers interested in mortality outcomes of enlistees may use either the CenSoc Enlistment-DMF file or the CenSoc Enlistment-Numident file. Both these datasets link enlistees to year, month, and age of death as recorded by the Social Security Administration. The CenSoc Enlistment-DMF file may be advantageous to some researchers because it contains more years of mortality data (1975-2005). In contrast, the CenSoc Enlistment-Numident covers only deaths in years 1988-2005. However, the CenSoc Enlistment-Numident has more death records per year on average, and allows researchers to access additional variables from Social Security Numident records, such as location of death. Where available, both these mortality datasets also contain links to the 1940 Census, allowing researchers to access even more information about certain enlistees. However, 1940 Census links are available for fewer than half of records in each mortality dataset. Overall, mortality researchers looking to capture the widest window of mortality data should consider using the CenSoc Enlistment-DMF file, while researchers wanting more covariates from mortality records may consider using the CenSoc Enlistment-Numident file.

6 Conclusion

In this paper, we introduce the CenSoc WWII Army Enlistment Dataset, a large source of cleaned and harmonized data on over 9 million men and women who enlisted in the United States Army. These data contain a vast amount of geographic and socioeconomic information on army enlistees. We also publish body height and weight for over 5 million individuals, making this dataset a uniquely large source of anthropometric data. Linked datasets, such as linked enlistment/CenSoc mortality data, open up many further research possibilities, and researchers may perform their own nominal linkages between enlistment records and other data. These data will allow researchers to gain new insights into the life course and morality outcomes of World War II service members.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature* 59(3).

Clever, Molly and David R. Segal. 2012. "After Conscription – Implications for the Armed Forces." *Sicherheit und Frieden* 30(1):9–18.
URL: *https://www.jstor.org/stable/24233116*

Electronic Army Serial Number Merged File, ca. 1938 - 1946. 2002. National Archives at College Park, College Park, MD.
URL: *https://catalog.archives.gov/id/1263923*

Ferrie, Joseph P., Karen Rolf and Werner Troesken. 2012. "Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence test scores among World War Two U.S. Army enlistees." *Economics and Human Biology* 10:98–111.

Goldstein, Joshua R., Monica Alexander, Casey Breen, Andrea Miranda González, Felipe Menares, Maria Osborne, Mallika Snyder and Ugur Yildirim. 2023. "Berkeley Unified Numident Mortality Dataset (BUNMD).".
URL: *https://doi.org/10.7910/DVN/TTWNK8*

Hull, Theodore J. 2006. "The World War II Army Enlistment Records File and Access to Archival Databases." *Prologue Magazine* 38(1).
URL: *https://www.archives.gov/publications/prologue/2006/spring/aad-ww2.html*

Karpinos, Bernard D. 1958. "Height and Weight of Selective Service Registrants Processed for Military Service During Wolrd War II." *Human Biology* 30(4):292–321.

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