CenSoc Methods Protocol V3

Monica Alexander, Joshua R. Goldstein, Casey Breen,
Ugur Yildirim, Maria Osborne

19 October, 2023

Introduction

The CenSoc project links the 1940 Census to mortality records, providing researchers with a new data resource for studying mortality. This methods protocol presents the procedure for constructing the CenSoc datasets.

The CenSoc Project has created three public-use datasets:

- The CenSoc-DMF dataset links the 1940 census to the Death Master File, a collection of over 83 million death records reported to the Social Security Administration. This matched file includes only men, as surname changes during marriage for women present challenges for accurate record linking. Our linking strategy relies on first name, last name, and year of birth. We use the ABE method developed by Abramitzky, Boustan, and Eriksson (2012, 2014, 2017). We limit links to those established with a conservative version of the matching algorithm, as described below. We weight to NCHS mortality totals by year, age, sex, race, and birthplace.

- The CenSoc-Numident dataset links the 1940 census to the National Archives' public release of the Social Security Numident file ("NARA Numident"). Our linking strategy relies on first name, last name, year of birth, and place of birth and uses the ABE method. To link unmarried women, we use father's last name as a proxy for women's maiden name. We limit links to those established with a conservative version of the matching algorithm, as described below. We weight to NCHS mortality totals by year, age, sex, race, and birthplace.

- The BUNMD is a cleaned and harmonized version of the NARA Numident file. The BUNMD is a single file comprised of the most informative parts of the 60+ application, claim, and death files released by the National Archives. All records are linked by Social Security Number. Variables of interest include race, place of birth, state in which the Social Security card was applied for, and ZIP code of residence at the time of death.

	CenSoc-DMF	CenSoc-Numident	BUNMD
Males-only	✓		
All-Gender		✓	✓
Link to 1940 Census	✓	✓	
SS Application Covariates		✓	✓
High Death Coverage	1975-2005	1988-2005	1988-2005
Size	~4.7 Million	~7.0 Million	~49.5 Million

R package

The CenSoc project has an accompanying internal R package, `censocdev`, which contains the code to construct the CenSoc data products. The package is available on GitHub: https://github.com/caseybreen/censocdev.

In addition, we are making our R implementation of the ABE data linkage algorithm publicly available. The package is available on GitHub: (https://github.com/uguryi/abeR).

Source Data

1940 Full-Count Census

The United States Census of 1940, taken on April 1st, 1940, collected demographic and socioeconomic information on 132,164,569 people. The 1940 Census records were released by the U.S. National Archives on April 2, 2012. The original records were digitized by Ancestry.com and are available through the Minnesota Population Center (MPC). The MPC provides a public, de-identified version of the complete count census as part of the IPUMS-USA project. However, names and other identifying information are not available in this public file. A secure version of files—with names and street addresses—must be accessed through a secure computing environment.

Version: Minnesota Population Center and Ancestry.com. IPUMS Restricted Complete Count Data: Version 3.0 [1940 Full Count Census]. Minneapolis: University of Minnesota, 2022.

Berkeley Unified Numident Mortality Database

The Berkeley Unified Numident Mortality Database, a cleaned and harmonized version of the National Archive and Administration's public release of the Social Security Numident File ("NARA Numident"), is a micro-level dataset with 49,459,293 death records. It includes information on name, race, sex, birthplace, ZIP code of residence at the time of death, and administrative variables, such as a person's age when they submitted their first Social Security application and their total number of Social Security applications. The death coverage is nearly complete for deaths for persons age 65+ for the window of 1988-2005.

Social Security Death Master File

The Social Security Death Master File (DMF) is a micro-level dataset with 85,822,194 death records. The DMF has variables reporting first and last name, middle initial, social security number, date of birth, and date of death, with nearly complete death coverage for persons age 65+ for the window of 1975-2005. A copy of the DMF was obtained from the UC Berkeley Data Lab in 2011.

Data Preparation

1940 Census

We performed the following processing steps on the 1940 Census:

1. Split the large dataset into a series of small datasets based on birthplace (using `split-by-bpl.R`). This makes the later matching process more manageable (described below) by allowing us to first match on small datasets and then simply concatenate these at the end.

2. Clean the first and last names in each of the small datasets in line with the name cleaning part of the ABE method (using `clean-names.R`). This procedure involves a series of cleaning steps, including removing common titles (e.g., Dr.) from names, replacing nicknames with their standard versions (e.g., Billy to William), removing non-alphabetic characters, and converting all names to lowercase.

Social Security Death Master File

We performed the following processing steps on the Social Security Death Master File:

1. Split date of birth and date of death to get day, month and year of birth and death (using `load-dmf-deaths.R`).

2. Clean the first and last names in each of the small datasets in line with the name cleaning part of the ABE method (using `clean-names.R`). This procedure involves a series of cleaning steps, including removing common titles (e.g., Dr.) from names, replacing nicknames with their standard versions (e.g., Billy to William), removing non-alphabetic characters, and converting all names to lowercase.

Berkeley Unified Numident Mortality Database (BUNMD)

We performed the following processing steps on the Berkeley Unified Numident Mortality Database:

1. Split the large dataset into a series of small datasets based on birthplace (using `split-by-bpl.R`). This makes the later matching process more manageable (described below) by allowing us to first match on small datasets and then simply concatenate these at the end.

2. Clean the first and last names in each of the small datasets in line with the name cleaning part of the ABE method (using `clean-names.R`). This procedure involves a series of cleaning steps, including removing common titles (e.g., Dr.) from names, replacing nicknames with their standard versions (e.g., Billy to William), removing non-alphabetic characters, and converting all names to lowercase.

Match Method

CenSoc-DMF: Match Method

We match the two datasets on first name, last name, and birth year using the ABE method. The ABE method requires exact matches on first name and last name while allowing for up to ±2 years of difference in birth year. Both "standard" and "conservative" versions of the matches are produced. Standard matches are all possible matches established using names and birth year, while conservative matches requires names to be unique ±2 years within and across datasets. We retain only conservative matches in the published CenSoc-DMF V3.0 dataset. For more information on the ABE method, see Abramitzky, Boustan, and Eriksson (2012, 2014, 2017). The specific steps are:

1. Load in the cleaned 1940 Census and Death Master Files datasets.
2. Match the two datasets using the ABE method (as implemented in `match-records.R`).
3. Restrict to deaths occurring 1975-2005 and conservative matches.

CenSoc-Numident: Match Method

We match the two datasets using the ABE method. For men and ever-married women, we match on first name, last name, birth year, and birthplace. For never-married women, we match on first name, father's last name, birth year, and birthplace. We perform three separate matches:

Match men:

1. For each birthplace that exists in both the 1940 Census and the BUNMD:

   a. Load in the cleaned Census and BUNMD datasets corresponding to that birthplace.
   b. Restrict Census and BUNMD datasets to males-only.
   c. Match the two datasets using the ABE method (as implemented in `match-records.R`).

2. Concatenate the resulting birthplace-specific datasets into a single dataset.

Match women ever-married in 1940:

1. For each birthplace that exists in both Census and BUNMD:

   a. Load in the cleaned Census and BUNMD datasets corresponding to that birthplace.
   b. Restrict Census to ever-married women and BUNMD to women.
   c. Match the two datasets using the ABE method (as implemented in `match-records.R`).

2. Concatenate the resulting birthplace-specific datasets into a single dataset.

Match women never-married in 1940:

1. For each birthplace that exists in both Census and BUNMD:
    a. Load in the cleaned Census and BUNMD datasets corresponding to that birthplace.
    b. Restrict Census to never-married women.
    c. Restrict BUNMD to women not already matched in ever-married women match.
    d. Match the two datasets using the ABE method (as implemented in `match-records.R`).
2. Concatenate the resulting birthplace-specific datasets into a single dataset.

We combine the matches from the male match, the ever-married women match, and the never-married women match to construct the CenSoc linked dataset. The final dataset is restricted to deaths occurring 1988-2005 and conservative matches within each relevent birthplace/sex/marital status group.

Weights

BUNMD Samples and Weights

We created two BUNMD samples with high death coverage. Sample 1 includes persons (1) born between 1900-1940 (2) dying between 1988-2005 and (3) a recorded sex. Sample 2 is the subset of Sample 1 records with complete information for birthplace, and race. For each sample, we constructed inverse inclusion-probability weights to the Human Mortality Database (HMD) on age at death, year of birth, year of death, and sex. We broke the sample into cells cross-classified by year of birth, year of death, age at death, and sex. We weighted each cell to the HMD "Deaths by Lexis triangles" totals. This allows aggregation to HMD totals by period or cohort.

$$W_j = \frac{\text{HMD deaths in cell j}}{\text{CenSoc deaths in cell j}} \tag{1}$$

CenSoc-Numident Weights

We constructed post-stratification weights to National Center for Health Statistics (NCHS) mortality totals for persons (1) dying between 1988-2005, and (2) dying between ages 65-100. We broke the CenSoc-Numident into cells cross-classified by year of death, age at death, sex, race (Black, White, or Other), and birthplace. For records where this process was inappropriate or impossible, we constructed alternative weights. Finally, weights were adjusted slightly in order to calibrate to population totals and eliminate extremely high weights. The detailed process for calculating weights is described in our technical docmentation.

CenSoc-DMF Weights

We constructed post-stratification weights to National Censter for Health Statistics (NCHS) mortality totals for people (1) dying between 1975-2005, and (2) dying between ages 65-100. We broke the CenSoc-DMF into cells cross-classified by year of death, age at death, sex, race (Black, White, or Other), and birthplace. For records where this process was inappropriate or impossible, we constructed alternative weights. Finally, weights were adjusted slightly in order to calibrate to population totals and eliminate extremely high weights. The detailed process for calculating weights is described in our technical docmentation.