Obtaining the full count 1940 census from IPUMS to use with CenSoc

Monica Alexander

March 2018

Contents

1 Introduction

The CenSoc dataset contains information about individaul's birth date (and implied at the 1940 census), death date and state of residence in 1940. The dataset does not contain any other information on the characteristics of the individuals; however, by using the unique identifier information contained in the CenSoc dataset, the data can be merged back to the 1940 census available on the IPUMS-USA website.

This document briefly outlines how to obtain data from the 1940 full count census, which is publicly available through IPUMS. Once obtained, the IPUMS data can be merged with the CenSoc dataset. This step is explained in the final section.

If you already have an IPUMS account and are familar with obtaining data from IPUMS, it may be more useful to head straight to the last section for details on how to merge. However, please note the section on 'Variables required for the CenSoc merge'.

2 Registering for IPUMS

The Integrated Public Use Microdata Series (IPUMS) is a collection of individual-level data sources. IPUMS consists of microdata samples from censuses and surveys, in both the USA and internationally. IPUMS is housed at the Minnesota Population Center at the University of Minnesota.

The IPUMS-USA data collection is free to use, but researchers must register and agree to terms and conditions. All you need to do to register is to fill out the form here:

https://uma.pop.umn.edu/usa/user/new?return_url=https%3A%2F%2Fusa.ipums.org%2Fusa-action%2Fmenu

(Note: this link was accessed in March 2018. If it is broken, go to `https://usa.ipums.org/usa/`, navigate to 'Login' and then 'Create an account').

3 Choosing variables from the 1940 full count census

Once you have an account, you are ready to select and download data from the 1940 full count census. To get started, go to `https://usa.ipums.org/usa/` and under 'CREATE YOUR CUSTOM DATA SET' click the 'GET DATA' button.

3.1 Select data sample

To restrict data extraction to the 1940 full count:

- Click the 'SELECT SAMPLES BUTTON'. This will take you to a page with all possible census and ACS data that are available.
- Uncheck the 'Default sample from each year' box
- Click the 'USA FULL COUNT' tab
- Check the 1940 100% box
- Click 'SUBMIT SAMPLE SELECTIONS' button

This will take you back to the variable selection page.

3.2 Variables required for CenSoc merge

There are several technical variables that are required for the CenSoc merge. These are:

- `SERIAL40`: 1940 census serial number
- `PERNUM`: person number in household
- `STATEFIP`: US state code

Together, these variables make up a unique identifier that allows correspondence across the two datasets. Additionally, the `SEX` variable is needed in order to restrict the sample to be males only.
To obtain these variables:

1. From the HOUSEHOLD drop down, select 'HISTORICAL TECHNICAL'. Then click on the cross next to `SERIAL40` to select it.
2. From the HOUSEHOLD drop down, select 'GEOGRAPHIC. Then click on the cross next to `STATEFIP` to select it.
3. From the PERSON drop down, select 'TECHNICAL'. Note that `PERNUM` should be `[preselected]`, but you can click on the cross next to it just to make sure.
4. From the PERSON drop down, select 'DEMOGRAPHIC'. Then click on the cross next to `SEX`.

Additionally, you may want to select `AGE` from PERSONAL → DEMOGRAPHIC. This is not required, but this variable should be equal to the `census_age` variable in CenSoc.

3.3 Other variables of interest

Now that you have variables required to merge the two datasets, you can choose other variables that you may be interested in studying mortality differences by. For example, you could select `RACE`, which is under PERSONAL → RACE, ETHNICITY, AND NATIVITY.

NOTE: the size of full count census data files is quite large, so you may want to restrict the addition of extra variables accordingly. For example, a file (containing males only) with `RACE` as an additional variable is around 6.6GB.

3.4 Restrict to males only

The CenSoc dataset only contains males, and given the large file size of the full count census, it is wise to restrict the data to only include males before you download it. To do this:

1. Click 'VIEW CART'
2. Check that all the variables you want are selected. Then click 'CREATE DATA EXTRACT'
3. Under OPTIONS, click the 'SELECT CASES' button
4. Check the box next to SEX
5. Chcek the box next to '1 Males' and click SUBMIT

4 Downloading the IPUMS file

To work with IPUMS data in `R`, it is usually easiest to download the data as a CSV file. To do this, on the EXTRACT REQUEST page, next to 'DATA FORMAT' click Change, select 'Comma delimited (.csv)' and click the submit button. Note that you can work with other formats in `R` as well, but CSV is generally the easiest. The only downside is that variable values are still listed in code format, and need to be converted to meaningful values using the codebook.

Once you are happy with your dataset, click the 'SUBMIT EXTRACT' button. Because it is full count data, you will need agree to special usage terms. Click Ok, and the dataset will begin to be extracted.

Given the size of the file, the processing may take a while. Once the file is ready, you will receive an email from IPUMS, and you can follow the links to download the resulting dataset.

5 Merging IPUMS with CenSoc

Once the IPUMS census dataset is obtained, it can easily be merged with the CenSoc dataset. An example `R` file can be found here: `https://github.com/MJAlexander/censoc/blob/master/ipums_merge_example.R`.

The idea is to create a unqiue ID by concatenating `SERIAL40`, `PERNUM` and `STATEFIP`. The three codes should be concatenated with some sort of separator (e.g. and underscore). The two datasets are then merged based on this unique ID.[1] This can be done using base `R`:

```
merged_df <- merge(censoc, census, by = "unique_d")
```

or using the tidyverse/dplyr functionality:

```
merged_df <- censoc %>%
  left_join(census, by = "unique_id")
```

or using the `data.table` package (which is probably the fastest):

```
merged_df <- censoc[census, nomatch=0]
```

[1] Note: there is no way of matching individuals who lived in residences for 60 or more people across the CenSoc/IPUMS census datasets. The combination of SERIAL40, PERNUM and STATEFIP is not unique.