

# Numident Supplementary Geography File Codebook

Page	Variable	Label
2	histid	Historical Unique Identifier
3	birth_string_uncleaned	Place of birth, uncleaned string
4	birth_gnis_code	Place of birth, GNIS code
5	birth_city	Place of birth, city
6	birth_county	Place of birth, county
7	birth_region	Place of birth, census region
8	death_zip	Place of death, 5-digit ZIP code
9	death_city	Place of death, city
10	death_county_fips	Place of death, FIPS county codes
11	death_county	Place of death, county
12	death_state	Place of death, state
13	death_region	Place of death, census region
14	death_country	Place of death, country
15	death_ruc1993	place of death, county rural-urban continuum

## Summary and Methodology

The Numident Supplementary Geography File (N = 7,044,519) provides a set of supplementary geography variables reporting place of birth and death for individuals in the Numident. This file can be linked onto the Numident at the individual-level using HISTID.

**Place of death variables:** To construct the place of death variables, we use the ZIP code of last residence available in the Numident Death record. It is likely that this reflects the ZIP code where an individual last lived. We map these ZIP codes onto city, county, state, census region, country, and rural urban continuum codes using a database from the United States Postal Service (USPS) Link. For ZIP codes that have been decommissioned, we use a secondary database from UnitedStatesZipCodes.org.

**Place of birth variables:** To construct the place of birth variables, we use the uncleaned 12-character city/county of birth string from the Numident Application records. These strings are uncleaned and contain misspellings and other inconsistencies. We mapped these uncleaned strings onto Geographic Names Information System (GNIS) codes using the crosswalk developed for this paper:

Black, Dan A., Seth G. Sanders, Evan J. Taylor, and Lowell J. Taylor. 2015. "The Impact of the Great Migration on Mortality of African Americans: Evidence from the Deep South." *American Economic Review*. 105(2):477–503. doi: 10.1257/aer.20120642.

To construct other place of birth variable, we mapped the GNIS codes onto city, county, state, and census regions using a database from the U.S. Board on Geographic Names Link. **Note that the District of Columbia and birth locations outside the United States are not mapped to GNIS codes.** Please cite Black et al. (2015) if you are using any of the birthplace geography variables in the file.

# HISTID

**Label:** Historical Unique Identifier

**Description:** HISTID is a unique individual-level identifier. This variable uniquely identifies all records in the dataset and can be used to merge onto the CenSoc-Numident mortality dataset.

# birth\_string\_uncleaned

**Label:** Place of birth, uncleaned city/county string

**Description:** birth\_string\_uncleaned reports the 12-character city/county of birth string from the Numident Application records. This variable is uncleaned and unprocessed, and any many contain spelling errors or inconsistencies.

NA values are due to birth cities not being inputted in social security records.

# birth\_gnis\_code

**Label:** Place of birth, GNIS code

**Description:** birth\_gnis\_code is a numeric variable that reports a person's Geographic Names Information System (GNIS) code for their place of birth. Each GNIS code maps onto physical locations located in the U.S., including longitude and latitude coordinates, physical feature names, counties, and states.

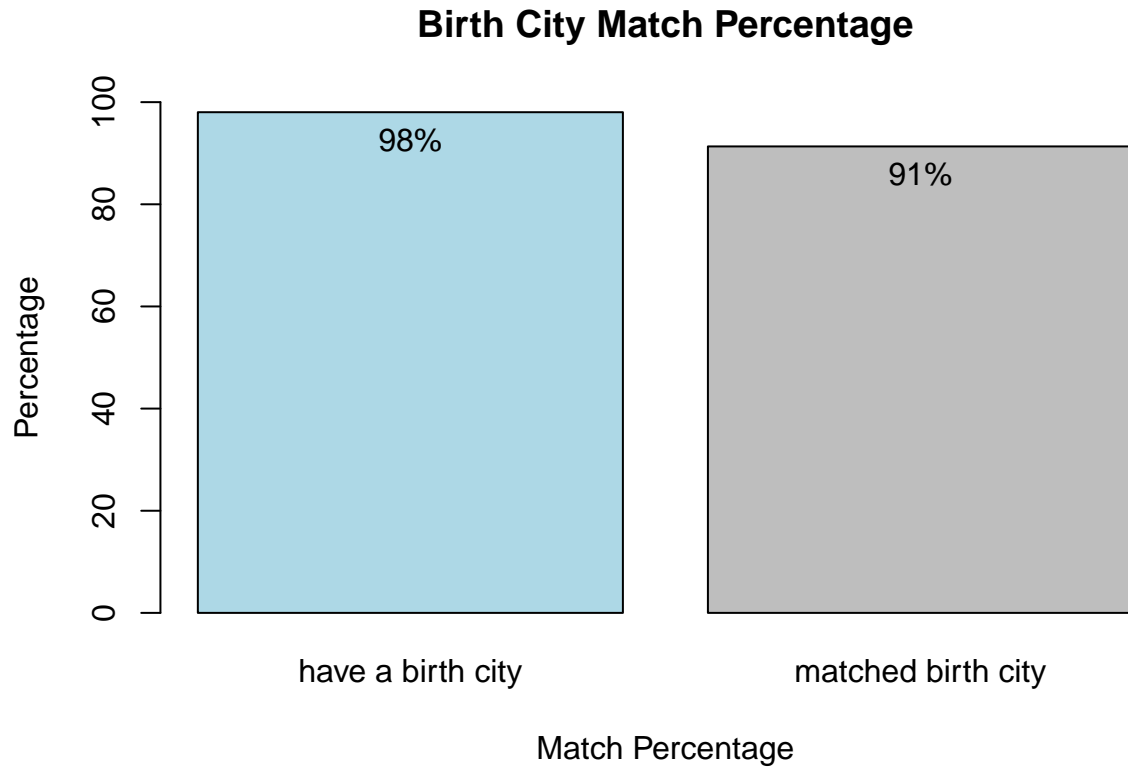
NA values are due to an inability to match uncleaned birth city values with the GNIS crosswalk or a lack of birth city on social security records.

For more information on GNIS codes, please see <https://www.usgs.gov/faqs/what-geographic-names-information-system-gnis>.

# birth\_city

**Label:** Place of birth, city

**Description:** birth\_city is a character variable reporting a person's city of birth. NA values are due to an inability to match uncleaned birth city values with the GNIS crosswalk or a lack of birth city on social security records.



# birth\_\_county

**Label:** Place of birth, county

**Description:** birth\_\_county is a character variable reporting a person's county of birth.

NA values are due to an inability to match uncleaned birth city values with the GNIS crosswalk or a lack of birth city on social security records.

# birth\_\_region

**Label:** Place of birth, census region

**Description:** birth\_\_region is a character variable reporting a person's census region of birth. For more information on Census Regions, please see <https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html>.

NA values are due to an inability to match uncleaned birth city values with the GNIS crosswalk, birth city being outside of a census region (in a different country for example), or a birth city not being entered in social security records.

Births by Census Region

Region	n	freq %
midwest	2405726	34.2
northeast	1773948	25.2
south	2078005	29.5
west	597450	8.5
NA	188302	2.7

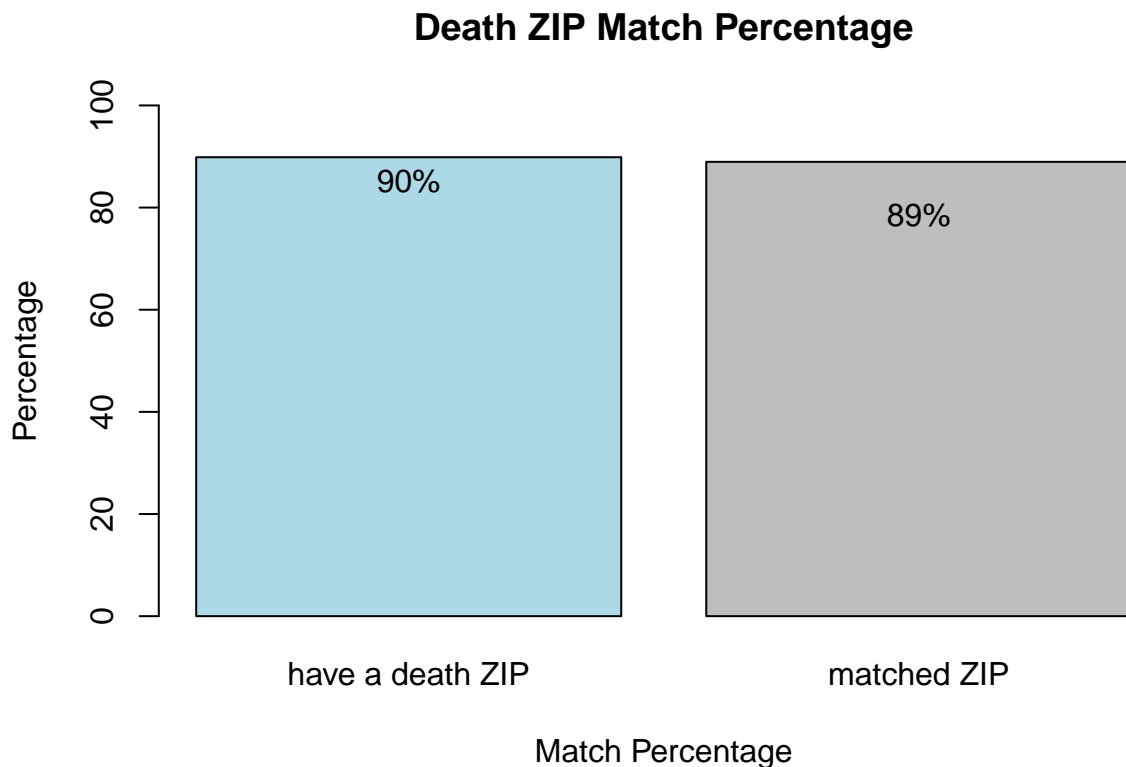
# death\_zip

**Label:** Place of death, ZIP code

**Description:** death\_zip is a numeric variable that represents the 5-digit ZIP code of the last residence. It is obtained from the Numident Death record, which is a record maintained by the Social Security Administration. However, it's important to note that the ZIP codes in this variable include non-identifiable codes such as "XX768" and "00000".

We note that ZIP codes are primarily used for USPS mail delivery purposes, not as geographic units. For a more detailed description of ZIP codes, please see: <https://faq.usps.com/s/article/ZIP-Code-The-Basics>.

NA values are due to ZIP code not being entered correctly in social security records or a person dying in a place where there is no USPS identifiable ZIP code.





# death\_city

**Label:** Place of death, city

**Description:** death\_city is a character variable reporting a person's city of death, as sourced from the ZIP code of their residence at time of death. For ZIP codes that cross city lines, we report the city where the primary post office for that ZIP code is located.

NA values are due to ZIP code not being entered correctly in social security records, a person dying in a place with no ZIP codes, or a ZIP code that could not be matched against USPS records.

# death\_\_county\_\_fips

**Label:** Place of death, county FIPS code

**Description:** death\_fips is a numeric variable reporting a person's county FIPS code of death. NA values are due to ZIP code not being entered correctly in social security records, a ZIP code that could not be matched against USPS records, a person dying in a place with no ZIP code, or a county not being matched with a FIPS code.

For a more detailed description of county FIPS codes, please see: <https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt#:~:text=FIPS%20codes%20are%20numbers%20which,to%20which%20the%20county%20belongs>.

Although FIPS codes rarely change, one major change occurred after Dade County, FL became Miami-Dade and a new FIPS code was assigned. If you're using FIPS code data that was produced after 1997, you'll want to manually change Dade County's FIPS code in order for it to match. Below is an example of how you could use these FIPS codes if you wanted to incorporate county-level education attainment (<https://www.ers.usda.gov/webdocs/DataFiles/48747/Education.xls>).

```
#Hypothetical education data
education <- read_excel("/data/censoc/workspace/geography_variables/Education.xlsx")

#we rename FIPS to death_fips so that it can match
colnames(education)[1] <- "death_fips"

#recode Miami-Dade's FIPS to Dade County's FIPS
education <- education %>%
  mutate(death_fips = ifelse(death_fips == "12086", "12025", death_fips)) %>%
  mutate(death_fips = as.character(death_fips))

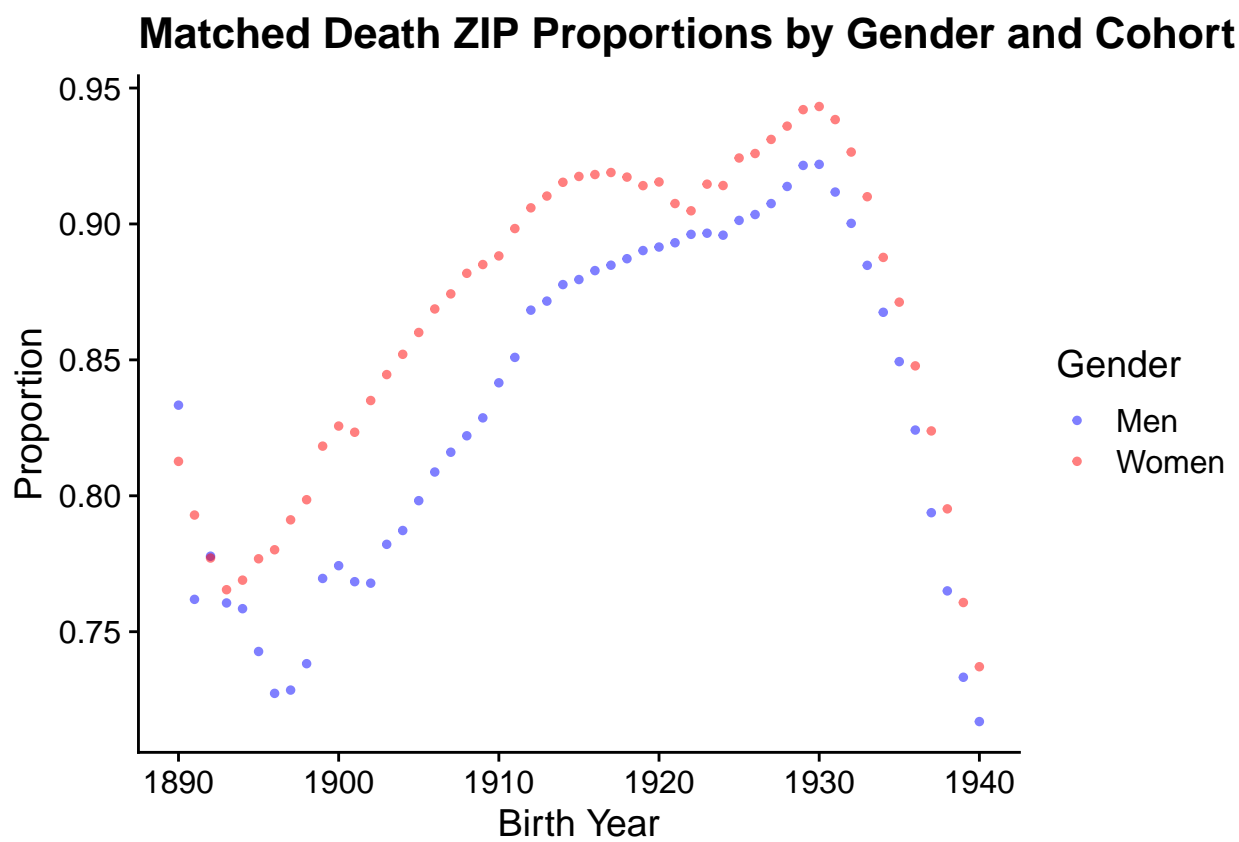
#merge onto Numident
numident_merged <- merge(numident_merged, education, by = "death_fips", all.x = TRUE)
```

# death\_\_county

**Label:** Place of death, county

**Description:** death\_\_county is a character variable reporting a person's county of death, as sourced from the ZIP code of their residence at time of death.

NA values are due to ZIP code not being entered correctly in social security records or a ZIP code that could not be matched against USPS records.



# death\_\_state

**Label:** Place of death, state

**Description:** death\_\_state is a character variable reporting a person's state of death, as sourced from the ZIP code of their residence at time of death.

NA values are due to ZIP code not being entered correctly in social security records, the person died outside of the US, or a ZIP code that could not be matched against USPS records.

# death\_\_region

**Label:** Place of death, census region

**Description:** death\_\_region is a character variable reporting a person's census region of death, as sourced from the ZIP code of their residence at time of death.

NA values are due to ZIP code not being entered correctly in social security records, the person died outside of one of the four census regions (for example in a different country), or a ZIP code that could not be matched against USPS records.

Deaths by Census Region

region	n	freq %
midwest	1538285	21.8
northeast	1301364	18.5
south	2057723	29.2
west	1368277	19.4
NA	777782	11.0

# death\_\_country

**Label:** Place of death, country

**Description:** death\_\_country is a character variable reporting a person's country of death, as sourced from the ZIP code of their residence at time of death. While the majority of ZIP codes are within the U.S., countries that host U.S. military bases on their territories may also have ZIP codes associated with them.

NA values are due to ZIP code not being entered correctly in social security records or a ZIP code that could not be matched against USPS records. People who died in a country where there are no ZIP codes identifiable by the USPS are assigned an NA.

# death\_ruc1993

**Label:** County of death, rural-urban continuum code

**Description:** death\_ruc1993 report the Rural-Urban continuum code for a person's county of death, as sourced from the United States Department of Agriculture (USDA). The Urban-Rural Continuum Codes are a classification system developed to categorize U.S. counties based on their degree of urbanization and adjacency to metropolitan areas. The system consists of ten codes, ranging from 0 to 9, where lower codes represent more urbanized counties and higher codes represent more rural areas.

NA values are due to ZIP code not being entered correctly in social security records, a ZIP code that could not be matched against USPS records, or a county not being matched with a FIPS code.

For more information: <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx>

## Rural-Urban Deaths

code	description	n	freq %
0	Central counties of metro areas of 1 million population or more	2448637	34.8
1	Fringe counties of metro areas of 1 million population or more	239373	3.4
2	Counties in metro areas of 250,000 to 1 million population	1410839	20.0
3	Counties in metro areas of fewer than 250,000 population	556812	7.9
4	Urban population of 20,000 or more, adjacent to a metro area	291965	4.1
5	Urban population of 20,000 or more, not adjacent to a metro area	172435	2.4
6	Urban population of 2,500 to 19,999, adjacent to a metro area	511677	7.3
7	Urban population of 2,500 to 19,999, not adjacent to a metro area	422889	6.0
8	Rural or fewer than 2,500 urban population, adjacent to a metro area	80740	1.1
9	Rural or fewer than 2,500 urban population, not adjacent to a metro area	128986	1.8
NA		NA	11.1