Author: Kien Hoang-Le

University of Nottingham

Created date: 05 July 2023

--------------------------------------------------------

**Notes on Data Wrangling Process**

This is a diary on the data wrangling process on the project: “Mortgage Pricing and Toxic Release”. The aim of the process is to create a joined panel dataset from HMDA data and Toxic Release Inventory Dataset.

Along with this document is another README text file, which is a quick guideline on the coding files (the purpose of each file and the code order).

## **Data Sources:**

### **HMDA data:**

Home Mortgage Disclosure Act (HMDA) data, provided by the U.S. Consumer Financial Protection Bureau. The data collect information of approximately 8mil. ~ 9mil. mortgage loans.

There are two waves of data due to differences in collection method: 2007 – 2017 and 2018 – present.

Source: <https://ffiec.cfpb.gov/data-publication/2022>

### **TRI data**

Toxic Release Inventory (TRI) data, provided by U.S Environmental Protection Agency.

Source: <https://www.epa.gov/toxics-release-inventory-tri-program>

Data Dictionary: <https://1drv.ms/b/s!Agir4OAFCUXNgkBNdmAVDPh2jmFK?e=ucS14M>

### **FIPS Code dataset**

Source: <https://walker-data.com/tidycensus/reference/fips_codes.html#details-1>

## Notes on the process

**April 10, 2023: Cleaning the TRI dataset**

The TRI dataset has a problem with the naming convention, so it needs to be renamed all the column variables before use.

After renaming all of the variables, I used the following variables and criteria to remove the “bad” observations in TRI dataset:

* Filter on **total\_releases:** > 0

I also select a set of following variables to use:

**July 05, 2023: Cleaning the HMDA dataset**

The original HMDA dataset for each year is from 7 mil. to 10 mil. records. I use the following variables and criteria to remove the “bad” observations in HMDA datasets:

* Filter on **loan\_term:** ==30 years (360 months)
* Filter on **loan\_to\_value\_ratio**: from 0 to 100
* Filter on **rate\_spread**: from –4 to 4 (covering 99% data)
* Filter on **interest\_rate**: lower than 12%
* Filter on **income:** higher than 0 and under the 99 percentiles
* Filter on **age**: smaller than 200, to remove 8888 and 9999
* Filter on **property\_value:** under the 99 percentiles

This process reduces ~30% observations number.

**July 05, 2023: Joining HMDA and TRI dataset to create a linked panel based on year-FIPS**

Create a testing panel data. The result is a panel with 9164 observations, including 2364 FIPS code and quite balanced throughout the year:

year N

1: 2018 2291

2: 2019 2297

3: 2020 2287

4: 2021 2289

Comparison check: (2018 sample)

* TRI: 2370 FIPS codes (counties), lowest is zipcode
* HMDA: 3187 FIPS codes, lowest is census\_tracts < zipcode (population)
* Panel: 2291 FIPS codes
* Total US FIPS code: 3242

This means that most of the TRI dataset has a match in HMDA, **but not all.**

**1 county has 1 Treatment variable**

**Treatment: 3 levels**

* **High pollution**
* **Low pollution**
* **No pollution (control)** (near factory)!!
* **Pollution variable:** total\_releases

**[mean(carcinogen), by = county] \* total\_releases =>**

**Weighted total\_releases**

**N-to-1 match**

**1000 mortgage rate – 1 county total releases**

**Avg(1000 mortgage rate) – 1 county**