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:**

* 1. **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>

* 1. **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>

* 1. **FIPS Code dataset**

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

* 1. **US Census county-level**

Source: <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html>

## **Notes on the process**

* 1. **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
* Filter on **classification:** Remove dioxin chemical release places since it might be too toxic when compared with other places.
* The zip code 99686 (FIPS code 02261) is removed since it belongs to Valdez–Cordova Census Area. It was a census area located in the state of Alaska, United States. As of the 2010 census, the population was 9,636. It was part of the Unorganized Borough and therefore had no borough seat. On January 2, 2019, it was abolished and replaced by the Chugach Census Area and the Copper River Census Area
* **I collect the following variables in the TRI dataset:**

|  |  |
| --- | --- |
| **Varlist** | **Description** |
| frs\_id, state, state\_code, latitude, longitude, county, fips, year, year\_fips, | Geolocator and ID variables |
| Carcinogen | Binary |
| Classification | Binary |
| x5.1\_fugitive\_air | Continuous (taking log in county level) |
| x5.2\_stack\_air | Continuous (taking log in county level) |
| onsite\_release\_total | Continuous (taking log in county level) |
| offsite\_release\_total | Continuous (taking log in county level) |
| total\_releases | Continuous (taking log in county level) |

* 1. **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.

The next step is to encode the character type values and bin continuous variables. The details are as follow:

|  |  |  |
| --- | --- | --- |
| **Varname** | **Type** | **Binning** |
| income | Continuous variable | Decile |
| Loan\_to\_value\_ratio | Continuous variable | Decile (there are duplicated breaks so a small random error is added in each value) |
| Property\_value | Continuous variable | Decile |
| Loan-to-income | Loan amount/Income | Decile |
|  |  |  |
|  |  |  |
|  |  |  |

* 1. **US Census variable**

Adding county-level census data from the US Census data

|  |  |
| --- | --- |
| POP\_COU | 2020 Census total population of the County |
| HOU\_COU | 2020 Census total housing unit count of the County |
| ALAND\_COU | 2020 land area of the County (square meters) |
| POPDEN\_COU | 2020 population density of the County (square miles) |
| ALAND\_PCT\_URB | Percent of 2020 land within the County that is classified as Urban |

* 1. **Creating Treatment Variable**

Creating treatment variables based on the TRI dataset.

The new dataset is collapsed based on the fips-year variable.

**Treatment: there will be 3 levels of treatments**

* **High pollution**
* **Low pollution**
* **No pollution (control)** (near factory)!

**Two questions arise: (1) what is pollution? (2) how do we define high and low?**

First, pollution defines as having carcinogenic substances in the **total waste releases**. That means treatment groups have carcinogenic waste, and the control groups don’t (for 2 levels) or treatment groups have 2 levels of carc waste. The second type of pollution is the level of carcinogenic in **fugitive air.**

Second, high and low is currently defined by median.