**Documentation for HMDAHarmonizer**

Zach Babat, last updated 2/17/23

## 1. Summary

**i. Context:** The Home Mortgage Disclosure Act (HMDA) dataset is a publication overseen by the Federal Financial Institutions Examination Council (FFIEC). HMDA contains information from thousands of lenders about tens of millions of mortgages and mortgage applications each year. It is one of the best public resources for studying mortgage lending in the United States. HMDA is published in single-year datasets, and contains two main components:

1. Loan-level data, where each observation corresponds to a mortgage application, origination, or purchase, and contains rich data regarding the characteristics of the loan and the applicant. Each observation includes a numeric code to identify the lender that reported the observation.[[1]](#footnote-1)
2. A lender panel, where each observation is a bank that filed a HMDA report. This panel crosswalks from the numeric bank identifiers to information about the bank (e.g. name), enabling researchers to study the lending activity of specific banks.

**ii. Problem:** Researchers may seek to use the lender identifiers to study the lending behavior of individual banks over time - for example, in a regression using multiple years of HMDA data with fixed effects for individual banks. However, between the annual publications of the dataset, it is possible for the numeric code that identifies a lender to change. This makes it difficult to perform studies that require consistently identifying the same lender in multiple years of the dataset.

**iii. Key Contribution:** This publication is a multi-year identifier panel. For a given bank, we provide a unique code that is linked to the identifier representing that bank in HMDA each year. This solves the problem described above by providing one harmonized code to track individual banks in HMDA, even if a bank’s identifier in the HMDA data changes between years. Our crosswalk covers the 2010-2021 HMDA datasets.

**iv. Structure and Use:** This multi-year identifier panel is in wide format. Each observation corresponds to a lender, defined by the unique identifier *masterid*.[[2]](#footnote-2) Each observation also contains a series of variables called *concatid[yyyy]*. The *concatid* variable for each year contains the code used to identify the lender in HMDA in that year. Note that the values of *concatid* can vary between years for a given borrower - *masterid* groups together all the ways a bank is represented over time. To use HMDAHarmonizer, a researcher can simply merge our panel onto a given year of the HMDA loan-level data, using the *concatid* variable for that year as the merge key. After repeating this process for multiple years of the dataset, the researcher can track a given bank by its *masterid*.

**v. HMDAHarmonizer File and Replication:**

The file HMDAHarmonizer panel is contained in the file, hmda\_harmonizer\_panel.dta. To replicate this file, follow the instructions in “\_readme\_sources” to download the input files, and then execute the script hmda\_­harmonizer.do

## 2. Panel Structure and ID Variables

**i. Identifier types:** There are 3 types of lender ID variables used in this crosswalk:

1. **Agency Codes and Respondent IDs:** From 2010-2017, lenders in HMDA are identified by the concatenation of a one-digit code identifying the regulatory agency and a longer numeric code sourced from the bank’s regulatory institution.[[3]](#footnote-3) We will refer to the concatenation of these codes as “HMDA IDs” or “pre-2018 HMDA IDs”. When banks change regulators, their pre-2018 HMDA ID codes change – we discuss how often this occurs below.
2. **LEIs:** From 2018-present, lenders in HMDA are identified by Legal Entity Identifier (LEI) codes from the Global LEI Foundation.[[4]](#footnote-4) LEIs are generally stable between years.
3. **RSSDs:** RSSD codes are not used as the official lender identifiers in HMDA, though they are included in the HMDA lender panels in all years. RSSD codes are issued by the National Information Center (NIC), which maintains data on financial institutions for which “the Federal Reserve has a supervisory, regulatory, or research interest.”[[5]](#footnote-5) RSSDs are designed such that each financial institution in the NIC database receives a single, unique RSSD as an identifier for its entire life cycle, and RSSDs are never reused.[[6]](#footnote-6)

Though a bank can report under multiple pre-2018 HMDA IDs or LEIs over its life cycle, each HMDA ID/LEI is unique within each year.

**ii. *masterid* and Defining Banks:** In our panel, we primarily defer to the RSSD codification system to distinguish between banks, and accordingly we primarily identify banks using RSSD codes. Based on this principle, we do the following to construct the *masterid* variable that defines a bank in our dataset. *masterid* is linked to the time-series of *concatid[yyyy]* variables that identify the bank in each year’s lender panel.

1. When available, use RSSD to define *masterid*, and construct the *concatid[yyyy]* variables as the pre-2018 HMDA IDs or LEIs associated with that RSSD in each year.
2. If a lender exists only in the post-2017 data but does not have an RSSD, we use LEI to define the bank, and assign LEI as *masterid*. *concatid[yyyy]* variables will be populated with the same LEI in the years that there is a HMDA report associated with that LEI.
3. If a lender exists in the pre-2018 data but does not have an RSSD, we use pre-2018 HMDA ID to define the bank, and assign an ad-hoc alphanumeric code as *masterid*. *concatid[yyyy]* variables will be populated with the pre-2018 HMDA ID in the years that there is a HMDA report associated with that HMDA ID.[[7]](#footnote-7)
4. We also perform steps to ensure that banks without an RSSD, but which appear in both the pre-2018 and post-2018 datasets, are linked together with a *masterid* that bridges across coding schemes. That is, the single *masterid* is linked to a bank’s representation via agency codes and respondent IDs from 2010-2017, and its representation via LEIs from 2018-2021.

## 3. Example Use

TK – I’ll walk through the JPMC example I wrote up a few weeks ago.

## 4. Sources, ID Stability, and Methodology

Before providing an overview of the methodology, here we discuss the sources used to create the HMDAHarmonizer panel and additional information about the relationships between the different types of ID codes and banks.

### Sources

The following list describes each of the sources used to create the HMDAHarmonizer panel, and a brief overview of the content of each source:

**HMDA Lender Panels:** Described above, this is the primary source of information for the panel. HMDA lender panels contain information for a given bank (e.g. name, city, assets) and the ID code used to represent the bank in the loan-level data.

**The Avery File:** Maintained by Neil Bhutta, the Avery File contains information for every bank that has ever filed a HMDA report. The Avery File is a wide-format dataset where observations correspond to each individual ID code that appears in HMDA, and contains a rich set of information related to a given ID code in each year that code is associated with a HMDA report. We use the Avery File as an auxiliary source of information to help us track banks as they change HMDA identifiers. For a detailed description of the differences between HMDAHarmonizer and the Avery File, see section 6.

**National Information Center (NIC) Datasets:** Described above, the NIC is the governmental organization that originates RSSD codes. We use data from the NIC to help match RSSD codes to LEIs when needed, and to improve the accuracy of our procedure to link banks by RSSD.

**HMDA-to-LEI Crosswalk:** This is an official resource published by HMDA to enable researchers to match banks from their pre-2018 HMDA IDs to their post-2018 LEI codes. This dataset is not used in the code to generate the HMDAHarmonizer panel, but the script includes commented-out code to demonstrate that there are no additional pre-/post-2018 matches that we do not already achieve via other methods.

The precise datasets downloaded, and instructions for how to download them, are contained in the “Sources.docx” file in this replication package.

### ID Stability

This crosswalk is built on the principle that there is a 1-to-1 correspondence between banks and RSSD codes, which we believe is warranted given the NIC’s stated description and intention of the RSSD system.

This contrasts with the pre-2017 HMDA IDs. Of all the financial institutions defined by unique RSSDs on record in a HMDA report before 2018, roughly 15% of them are assigned more than 1 HMDA ID between 2010-2017. Thus, an analysis that uses HMDA ID alone would erroneously classify each of those banks as multiple unrelated entities.

Notably, the LEIs used in post-2017 HMDA are far more stable. Of the banks that file a HMDA report beginning in 2018 and that have an RSSD on record, there are only 5 RSSDs linked to more than 1 LEI, and only 20 LEIs linked to more than 1 RSSD. Thus in the procedure, we leverage the principle that banks with the same LEI are the same bank (and make adjustments in the rare cases when this seems untrue).

### Methodology

Below is an overview of the procedure to create the HMDAHarmonizer panel. A detailed walkthrough of this procedure is contained in section 7, including descriptions of checks to demonstrate the accuracy of measures we take to match together ID variables in different years. That walkthrough also refers to Appendix B, which contains detailed descriptions of the [X] cases where I manually recode RSSD information.

Our plan is to generate RSSD-based *masterids* to link the pre-2018 HMDA identifiers together, do the same for the LEIs in the post-2018 HMDA, then to merge the two epochs together using *masterid.*

Generating pre-2018 *masterids:*

1. Using RSSD as a merge key, execute a series of 1:1 merges to combine all of the lender panels from 2010-2017. Before each merge, save observations where RSSD is missing, “0”, or non-unique to a separate tempfile. The result of these merges is a wide-format panel where observations are uniquely identified using RSSD, and each observation has a series of variables containing the HMDA ID corresponding to that RSSD each year.
2. Resolve a very small set of (roughly 10) RSSDs that are duplicated within a given year (sometimes this is due to erroneous RSSD coding, in rare cases we believe a single RSSD is associated with multiple HMDA IDs within a given year).
3. Resolve observations where a HMDA ID is not associated with an RSSD code in a given year.
   1. Check if a given HMDA ID is associated with an RSSD in a different year
   2. Look for RSSD information in the Avery file
4. Of the HMDA ID codes that did not match to RSSDs, group together HMDA ID codes that appear in multiple years and assign them an ad-hoc *metaid*.
5. Generate *masterid*:
   1. For the rows containing a series of HMDA IDs we were able to match to an RSSD, *masterid* is that RSSD code.
   2. For the rows containing a series of identical HMDA IDs that did not match to an RSSD, *masterid* is the ad-hoc *metaid*.

Generating post-2018 *masterids:*

1. Append together the post-2018 banks, save observations with RSSDs that correspond to more than 1 LEI to a separate tempfile, then reshape wide by LEI. Save a separate tempfile of LEIs not associated with an RSSD. The result of this is a wide-format panel where observations are uniquely identified using RSSD, and each observation has a series of variables containing the LEI corresponding to that RSSD each year.
2. Resolve LEI codes not associated with an RSSD:
   1. Look for RSSD information the NIC files
   2. Look for RSSD information in the Avery file
   3. Go back to the lender panels, which contain information on the pre-2018 HMDA ID corresponding to a given LEI (if any), and look for the *masterid* associated with that pre-2018 HMDA ID code. (Note that this *masterid* could be sourced from a *metaid*, not just an RSSD).
3. Generate *masterid*:
   1. If able to match an LEI to an RSSD in the steps above, masterid is RSSD
   2. If only able to match an LEI to a previously established *metaid*, masterid is metaid
   3. If neither 1 nor 2, then masterid is LEI
4. Resolve 6 cases of where *masterid* is duplicated.
5. Merge together pre-2018 and post-2018 panels, using *masterid* as a merge key.
6. Resolve cases where an LEI matched to multiple RSSDs from above, and append these rows to the merged dataset from the previous step.

Now, we have a panel with unique observations identifying individual lenders using the “masterid” variable, and containing the codes that identify that bank’s loans in the loan-level data for each year from 2010-2021.

Additional modifications:

(Full description TK)

* 1. Identify donuts and look for switchers
  2. Add on pre-2017 lenders that aren’t in the crosswalks
  3. Add on post-2017 lenders that aren’t in the crosswalks

## 5. Edge Cases and Exceptions

Discuss:

Duplicated masterids (duplicated reporting, switchers)

Missing RSSDs

Ad-hoc masterids

## 6. FAQ

## 7. Step-by-step Summary

## 8. Sources

## Appendix A: Codebook

## Appendix B: Documentation for recoding data in the HMDA lender panels and for various anomalies

1. There are different versions of the loan-level data, e.g. the “Snapshot” vs. “Dynamic” loan-level datasets, the difference between these is not important for our purposes. [↑](#footnote-ref-1)
2. See Section 4 of this documentation - for 33 different masterid codes, masterid is duplicated in two observations. [↑](#footnote-ref-2)
3. [HMDA Documentation (cfpb.gov)](https://ffiec.cfpb.gov/documentation/2017/identifiers-faq/) [↑](#footnote-ref-3)
4. [HMDA Documentation (cfpb.gov)](https://ffiec.cfpb.gov/documentation/2022/identifiers-faq/) [↑](#footnote-ref-4)
5. [About - National Information Center (ffiec.gov)](https://www.ffiec.gov/npw/Home/About) [↑](#footnote-ref-5)
6. NIC Data Dictionary, p. 21 [↑](#footnote-ref-6)
7. We perform a manual check to ensure that pre-2017 HMDA IDs consistently identifies the same bank in these cases. Note also that while pre-2017 HMDA IDs can change, there do not seem to be cases when a HMDA ID is “abandoned” by one lender and “re-used” by a different lender. [↑](#footnote-ref-7)