**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)

See section 4 for discussion of how often a given bank experiences a change in each of these types of ID codes.

**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

## 4. ID Stability, Sources, and Methodology

[Before describing the procedure in more depth, need to make an additional note about how these identifiers work, describe some auxiliary sources that we bring in]

i. Note on Identifier Stability: This crosswalk is built on the principle that there is a 1-to-1 correspondence between banks and RSSD codes. This is in contrast to the pre-2017 HMDA IDs. Of all the financial institutions defined by unique RSSDs that file 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 post-2017 and 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.

ii. Sources

[Before describing the procedure, we need to talk about some additional data sources that we bring in]

Avery File

NIC Datasets

HMDA-to-LEI Crosswalk

Not actually used - talk about general purpose, we only include it to show that we don’t need it

i. General Methodology: [TK - I think loosely I can describe the procedure as follows:

Pre-2017:

Match banks up by RSSD

Find RSSDs for banks that don’t have them initially, and resolve one duplicate RSSD case

Match together banks still missing RSSDs when possible

Post-2017:

Match banks together by LEI, and in most of these cases they already have RSSDs, which handles the pre-post merge

Match LEIs to RSSDs using NIC

Match LEIs to RSSDs using the Avery file

Use info from the lender panels to match missing-RSSD banks onto HMDA IDs

Merge together pre- and post-2017 panels into a single, masterid-identified panel

Resolve cases when LEI is matched to multiple RSSDs

Cleanup/modifications:

Identify donuts and look for switchers

Add on pre-2017 lenders that aren’t in the crosswalks

Add on post-2017 lenders that aren’t in the crosswalks

## 5. Edge Cases and Exceptions

e. Exceptions:

Duplicated masterids (duplicated reporting, switchers)

Missing RSSDs

Ad-hoc masterids

## 6. Step-by-step Summary

## 7. 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)