**Documentation for HMDAHarmonizer**

TABLE OF CONTENTS:

1. Summary
2. Panel Structure and ID Variables
3. Illustration and Use Guide
4. Sources, ID Stability, and Methodology
5. Duplicated *masterid* observations
6. FAQ
7. Step-by-Step Description
8. Sources
9. Appendix A: Codebook
10. Appendix B: Example use code
11. Appendix C: Documentation for recoding data in the HMDA lender panels and for various anomalies

## 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, time-invariant code that is linked to the identifier for 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 combination 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 combination 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. Thus, we primarily identify banks using RSSD codes. Using 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*. (This ad hoc code is also stored as the variable *metaid*). *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, even if a bank lacks an RSSD, we ensure the single *masterid* is linked to the complete series of pre-2018 HMDA IDs and post-2018 LEIs that identify the bank in each year of HMDA from 2010-2021.

## 3. Illustration and Use Guide

**i. Illustration:** The following is an example of what we hope to offer with this crosswalk: the ability to track lenders as they change HMDA IDs.

Below is a screenshot of the Avery File, showing the row for HMDA ID 10000000008 (where 1 is regulatory code, and 0...08 is the HMDA lender code). The variables NAME[yy] are populated with the bank name for this HMDA ID in each year, and are blank in years when this ID did not file a report. NAME10 is populated, but the following years are empty. ***This row shows that lender 10000000008 corresponds to JPMorgan Chase in 2010, but we do not observe JPMC’s lending in subsequent years.***



Below is a screenshot of the HMDAHarmonizer panel, showing the row for masterid 852218 (in this case, masterid is derived from an RSSD code). Note that 852218 is the RSSD for JPMC, so***this row shows the different codes used to identify JPMC in the HMDA loan-level data in each year.*** Up until 2017, these are HMDA IDs, and beginning in 2018 these are LEIs. The variables concatid[yy] contain the correct code in each year.



Note that the series of *concatid[yyyy]* variables changes twice - once from 2010 to 2011, when HMDA ID changes, and once from 2017 to 2018, when all banks switch to LEIs. Below is a screenshot of the names associated with these HMDA IDs/LEIs in each year:



This illustrates the central utility of our crosswalk: even though the HMDA code associated with JPMC changes twice, we have a single identifier (masterid) that is linked to JPMC’s HMDA code in each year.

**ii. How to use HMDAHarmonizer:**

To use the HMDAHarmonizer panel, the user must merge the panel onto a given year’s loan-level data, using the corresponding *concatid[yyyy]* variable as the key in a one-to-many merge.

Recall that in years 2010-2017, lenders in HMDA are identified by the combination of a one-digit code identifying the regulatory agency and a longer numeric code sourced from the bank’s regulatory institution. Accordingly, for years 2010-2017, the *concatid[yyyy]* variable contains the agency code in the first digit, and the bank’s respondent ID in the remaining digits. It is critical for the user to split the *concatid[yyyy]* string into its these parts to perform a merge. See codebook for details and Appendix B for example code.

In years 2018-2021, the *concatid[yyyy]* variable contains a lender’s LEI. Banks are identified in the post-2018 data with this LEI, thus no further modification is needed to perform the merge beyond changing the variable name from *concatid[yyyy]* to *lei*.

After performing these merges, every observation of the loan-level data will be tagged with a *masterid*, which can be used to study the observations reported by the same bank in multiple years.

## 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.

**i. 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:* The NIC is the governmental organization that originates RSSD codes. The NIC datasets include crosswalks from RSSDs to LEIs. They also include data regarding bank mergers, closures, and other events in which a bank’s RSSD may change.

*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 described in the “Sources.docx” file in this replication package.

**ii. 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 rare cases when this appears untrue).

**iii. 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 41 cases where I either manually recode the relationship between a bank’s RSSD and its HMDA ID/LEI in a given year, or when I alter a bank’s *masterid* (described more at the end of this section).

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 where “masterid” is linked to *concatid* variables containing the codes that identify that bank’s loans in the loan-level data for each year from 2010-2021.

**Additional modifications:**

1. Look for banks with multiple RSSDs

In rare cases, there is evidence to suggest that a single institution is actually associated with multiple RSSD codes over its lifespan (see section 5). I perform a manual audit of HMDAHarmonizer to find instances of this occurring, and link together the two RSSDs with a single *masterid*. I find 15 pairs of RSSDs that I believe should be linked together with one *masterid* for each observation in the pair.

1. Add on lenders that are not in the crosswalks

Some lenders appear in the HMDA loan-level data, but not the HMDA lender panels. In the pre-2018 data, there are two years that each contain a single HMDA code that is not contained in the corresponding lender panel. In the step-by-step summary I discuss the procedure I use to match these banks to an already existing time-series of identifiers in HMDAHarmonizer.

In the post-2018 data, there are 138 LEI codes that appear in the loan-level data but not the lender panels. This replication package includes the file *banks\_not\_in\_lender\_panel,* which contains those LEI codes and the years they appear in the loan-level data. I use the NIC datasets and Avery files to match these LEIs to RSSDs when possible. I assign *masterid* using RSSD when possible, and assign *masterid* using LEI when RSSD is unavailable, then merge these banks onto the HMDAHarmonizer panel.

1. Incorporate data from the 2020 and 2021 Avery files

When the bulk of this code was written, the 2020 and 2021 Avery files were not yet published. At the end of the script, we check for cases when these Avery files contains RSSD-LEI pairings that do not match what is in our dataset. There are fewer than 100 such cases – see section 7 for details.

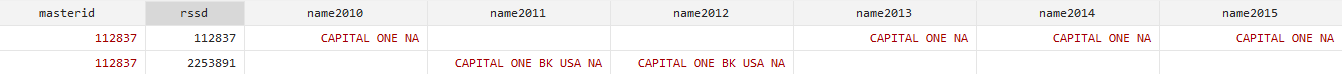
1. Add recoding flag

At the end of this procedure, there are 41 masterid-defined banks that experience some type of manual recoding. Whereas most of this procedure is based on ex-ante defined rules to try and match banks together, occasionally we intervene and either change the relationship between an RSSD and a HMDA ID/LEI, or we change the masterid associated with an RSSD (as in the “multiple RSSD” procedure described above and in section 5). To help researchers who would like to exclude banks that experience these ad hoc judgements from their analyses, we add a binary variable called *recoding\_flag*, which equals 1 when a given *masterid* was subject to a manual intervention. Users can simply exclude observations where *recoding*\_*flag* equals 1 from their analyses if they wish.

## 5. Duplicated *masterid* observations

**i. Duplicated *masterid* codes:** The HMDAHarmonizer panel is organized on the general principle that there is a 1-to-1 correspondence between RSSD codes and banks. In rare cases, there is reason to believe this is not the case, resulting in 33 *masterid* codes that appear in two different observations to group the multiple RSSDs associated with a single bank together.

**Type A – RSSD switchers:** Below are the two rows corresponding to the only RSSDs that ever appear in a HMDA lender panel with the words “Capital One” in the bank name. Note the abbreviation “NA,” which refers to “National Association.”



Looking only at the top row, it appears that Capital One performs mortgage lending in 2010, stops in 2011 and 2012, and 2013. Looking at the bottom row, it appears that a slightly differently titled Capital One performs mortgage lending only in 2011 and 2012, and then never again. This is highly implausible – it is far more likely that the national association for Capital One filed HMDA reports under different RSSD codes in different years.

Our solution to this is to keep two rows for Capital One, one row for each RSSD. However, the two rows are linked using one value of *masterid*, so a user can still track all the loan-level observations associated with Capital One in different years with a single identifier code.

These RSSD-switchers comprise 28 of the 33 *masterid* duplicates. Explicit details on the process to identify *masterid* switchers can be found in the step-by-step procedure in section 7.

**Type B – Duplicate filers:** There are 5 *masterid* duplicates not described in the section above. These correspond to RSSDs that are associated with multiple HMDA ID codes in a single year. After I perform background research on the banks listed in the lender panel with those HMDA ID codes, I conclude that both ID codes correspond to a single economically meaningful institution (the duplicates are related to some type of purchase or merger activity, leading to two filings associated with one RSSD).

**Note that regardless of these issues, the HMDA lender codes contained in the *concatid[yyyy]* are still unique within each year.**

**Note**

## 6. FAQ

TK

Questions to consider:

1. Why didn’t you use the more recent editions of the Avery file?
   1. A: I’ve run a check and I don’t need to – provide more detail after the check is complete

## 7. Step-by-step Description

**Note – in steps when RSSD recoding is described, the part of the .do file where that recoding occurs is marked with a comment giving the number of the corresponding step in this checklist (e.g. searching “2.a.i” in the .do file will take the user to where the recoding in 2.a.i occurs)**

1. Merge together HMDA lender panels, pre-2018

1. Looping merge – for each year of the HMDA lender panel datasets, do the following:
   1. Create a tempfile containing observations where RSSD is missing, “0”, or non-unique - we call these “problem banks”, each year’s problem banks get appended together (long format)
   2. Drop the observations in step i from the lender panel
   3. Merge together lender panels, without the problem banks, using RSSD as the merge key (merge 2011 onto 2010, merge 2012 onto 2010/2011, and so on)

*Now we have a wide-format “main panel,” with RSSD as an observation key, containing the HMDA IDs for each RSSD in each year*

2. Resolve “problem bank” cases where RSSD is listed as “0,” RSSD is missing, or RSSD is duplicated in a given year

1. Resolve observations with duplicate RSSDs
   1. Recode banks that have likely been assigned the wrong RSSD on the basis of name similarity to another bank (3 RSSDs)
   2. Do not recode cases where it is reasonably judged that the same RSSD reports under two HMDA IDs in a given year (4 RSSDs, no recoding)
   3. Reshape wide to prepare for merging back onto the “main panel” later, and save tempfile
2. Resolve observations with missing or “0” RSSDs using HMDA IDs - call both of these “missing” RSSDs
   1. Check HMDA IDs are stable for these banks
      1. With the data in long format, count the number of names associated with a HMDA ID
      2. For HMDA IDs that match to more than one name, confirm these are just cosmetic name changes (e.g. changes in punctuation, or changes in name documented on the bank website)
      3. Confirm no bank names match to more than one HMDA ID
   2. Match missing-RSSD observations to non-missing RSSD observations from the main panel
      1. Reshape the main panel to long-format and append onto missing RSSD observations
      2. Sort by HMDA ID and year, keep only HMDA IDs that have at least one missing RSSD observation and one non-missing RSSD observation
      3. For these banks, we now have missing-RSSD observations matched to non-missing-RSSD observations of those same banks from other years, using common HMDA IDs
      4. Confirm each HMDA ID matches to only one non-missing RSSD
      5. Recode cases where HMDA ID matches to more than one non-missing RSSD (1 RSSD)
      6. Reshape wide and perform an update merge of these banks onto the main panel
      7. Set aside another bank that seems to legitimately report under 2 RSSDs in a given year (1 RSSD, no recoding)
      8. Recode HMDA ID in 2010 for two banks with an identical name that I believe are assigned the wrong HMDA IDs (2 RSSDs)

*Now I have updated the main panel of RSSD-identified banks with HMDA IDs that were initially missing in certain years*

1. Of the missing RSSD observations that did not match to an RSSD in step 2b, attempt to find RSSD information in the Avery File
   1. Reload the list of missing RSSD observations, drop the observations that matched to an RSSD in the previous step
   2. Using HMDA ID as a key, merge on RSSD[year] variables from the Avery file
      1. We get two pairs of duplicated RSSD codes in 2010 - determine correct recodings to preserve unique RSSDs. Note that this is not recoding HMDA lender panel data, but rather correcting the results of the merge with the Avery file
   3. Using HMDA ID, match these banks to their names in the HMDA lender panels to double-check this method of matching banks to RSSDs is valid
      1. Recode the RSSD for one bank (1 RSSD)
   4. Perform an update merge of these banks onto the main panel

*Now I have once again updated the main panel of RSSD-identified banks with HMDA IDs that were initially missing in certain years*

1. Of the banks that still have not matched onto RSSDs, we assume these banks truly have no RSSD. Generate an ad hoc “metaid” variable and merge these onto the main panel
   1. Reload the list of missing RSSD observations, drop the observations that matched to an RSSD in the previous steps
   2. Confirm that the remaining HMDA IDs correspond to only one bank name, or that multiple names signify only cosmetic name changes (e.g. changes in punctuation, or changes in name documented on the bank website)
   3. Sort banks by HMDA ID, and generate a variable called “metaid” that equals the largest row number for each HMDA ID, plus the letter “A” in front (e.g. “A93” - the letter “A” distinguishes banks with this ad hoc ID from banks whose IDs are sourced elsewhere)
   4. Reshape to wide format, and append banks to the main panel (do not merge - recall the assumption that these banks do not match any banks identified with an RSSD)

*Now all “problem bank” observations are either in the main panel, or have been processed and set aside to be merged on in the next step*

1. Clean-up for pre-2017 banks
   1. Generate a variable called “masterid,” which takes the value of either the RSSD for a row or the metaid for a row (no row has both), so that we have one variable with the unique identifier for all banks
   2. Merge on the 4 duplicate RSSD banks from step 2a, plus the additional duplicate RSSD bank from step 2bii7

*Now all HMDA IDs contained in the 2010-2017 lender panels have been arranged in a crosswalk containing stable identifiers over time*

3. Merge on post-2017 data

1. Merge post-2017 lender panels together
   1. Append the 2018-2021 lender panels together (data in long format)
   2. Identify cases where more than one RSSD corresponds to an LEI, and save a tempfile of these observations
   3. Drop multi-RSSD LEIs from the dataset, and reshape wide
   4. Save a tempfile of missing RSSD observations (RSSD == -1)

*Now we have a wide-format post-2017 “main panel,” with RSSD as an observation key, containing LEIs that identify observations in HMDA for each RSSD in each year*

1. Identify missing-RSSD banks using the NIC dataset
   1. Append together “Attributes - Active” and “Attributes - Closed” datasets from the NIC
   2. Using LEI as a merge key, merge the NIC datasets onto the missing-RSSD banks
   3. Keep the banks that match to RSSDs, and perform an update merge of onto the main post-2017 panel (using LEI as a merge key)

*Now I have updated the main panel of post-2017 banks with RSSDs that some LEI codes were originally missing*

1. Identify missing-RSSD banks using the Avery file
   1. Reload the list of missing-RSSD banks, drop the observations that matched to an RSSD in the previous step
   2. Using LEI as a key, merge together 2018 and 2019 Avery files, and merge these onto the missing-RSSD banks
   3. Keep the banks that match to RSSDs, and perform an update merge of onto the main post-2017 panel (using LEI as a merge key)

*Now I have again updated the main panel of post-2017 banks with RSSDs that some LEI codes were originally missing*

1. Match missing-RSSD banks to pre-2017 banks by HMDA ID
   1. Reload the list of missing-RSSD banks, drop the observations that matched to RSSDs in the previous steps
   2. Merge the list of missing-RSSD banks back onto the post-2017 lender panels, which contain information on pre-2017 HMDA IDs. Our goal is to use that information to link post-2017 banks back to pre-2017 counterparts.
      1. In the script, I also include a commented-out block of code demonstrating that we gain no additional matches by using the official HMDA document, “ARID2017 to LEI Reference Table,” which can be found here: https://ffiec.cfpb.gov/documentation/2022/identifiers-faq/.
      2. I also demonstrate, lower down, that we gain no additional matches by looking for pre-2017 HMDA IDs in the Avery file.
   3. Keep the missing-RSSD banks that matched to valid pre-2017 HMDA IDs from the individual lender panels, match them onto their corresponding masterid codes from the pre-2017 dataset using pre-2017 HMDA ID as merge key.
      1. Prior to this, I discover that 12 post-2017 banks that found pre-2017 HMDA IDs do not match onto pre-2017 masterid codes. I discuss the resolution to this in Appendix B - 6 banks are matched back to masterid codes using pre-2017 HMDA IDs, and the other 6 are deemed invalid matches.

*Now, I have again updated the main panel of post-2017 banks, matching some post-2017 banks onto the “masterids” used to identify banks in the pre-2017 data. Note that in all of these cases, the source of this “masterid” is “metaid” - these banks are not identified with RSSD but rather our ad hoc ID codes.*

1. Merge post-2017 data onto the pre-2017 panel
   1. Generate masterid:
      1. If able to match an LEI to an RSSD in the steps 3a-3c above, masterid = RSSD
      2. If not 1, but able to match an LEI to a metaid in step 3d above, masterid = metaid
      3. If neither 1 nor 2, then masterid = LEI
2. Correcting 6 instances of duplicated masterids - see Appendix B
   * 1. In 4 cases, these are cosmetic changes related to LEI typos in single years
     2. The other 2 banks I believe are rare instances of 1 RSSD switching between LEIs in different years. I collapse each of these into 1 row per LEI. (2 RSSDs affected)
3. Using masterid as a merge key, merge the post-2017 ID panel onto the pre-2017 ID panel
   1. Clean-up steps including misc. reorderings, renamings, etc.
4. Resolve cases when the same LEI matched to multiple RSSDs (from 3aii).
   1. See Appendix B - I determine the correct RSSD to use as masterid for each bank (35 RSSDs) and merge onto the main panel from the step immediately above.

***Now we have 1 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.***

(Note there are rare cases in which individual masterids report to HMDA under two ID codes in a given year, as in step 2e, or switch between RSSDs/LEIs, as in step 3f.)

**Continued on next page**

4. Look for “RSSD Switchers” and “HMDA ID Donuts”

One could consider the panel “complete” at this point. In this section, I perform a manual audit of the HMDA Crosswalk to find cases when two different masterid’s actually both identify the same bank. When I am confident that this occurs, I tag both rows of the crosswalk with the same masterid. See Section 5 for further discussion of what leads to duplicated *masterid* codes.

In particular, I am trying to identify:

1) “Switchers” - this term refers to cases when an institution changes its RSSD over our observed timeframe. In such cases, we will need to link the distinct RSSDs with the same masterid.

2) “Donuts” - this term refers to cases when a bank is present in the HMDA data in one year, not present in a later year, then is present again in a year after that. Sometimes, a donut occurs because a bank truly is not represented in HMDA in a given year. Other times, it occurs when a bank temporarily files under a different RSSD.

1. Use NIC “transformations” dataset to look for banks that switch RSSDs
   1. Load in our wide-format HMDA ID crosswalk, as completed immediately above in step 3f
   2. Drop rows where masterid is duplicated, so we keep only banks that we have not already identified as having multiple RSSDs/LEIs
   3. Reshape long
   4. Grouping by HMDA ID, and count how many masterids are associated with each bank. Keep only banks with HMDA IDs that get mapped onto by more than one RSSD, these are RSSD-switcher candidates
   5. Using RSSD as a merge key, merge on the NIC “transformations” data
   6. See Appendix B - using the NIC dataset on RSSD transformations as a guide, I researched these banks where a single HMDA ID matches to multiple values of masterid. I judged which banks are likely to be the same institution reporting under different RSSD codes (which should be linked with the same masterid), and which are likely to be distinct institutions (which should keep their distinct masterid codes)

*At the end of this process, I reload the wide-format crosswalk. I identify 18 rows which I believe, in reality, correspond to 9 lenders that simply report with different RSSDs. I group these 18 rows into 9 pairs of rows, linked with the same masterid*

1. Identify HMDA “donuts,” where a given masterid-identified bank has gaps in its reporting
   1. Take the wide-format HMDA Crosswalk and reshape long
   2. Identify all masterids that file a HMDA report, experience a spell of non-reporting, and then report again later. These are the banks we call “donuts”.
   3. Keep the list of masterids that correspond to “donuts,” merge this list onto the HMDA Crosswalk, generate a “donut” variable to flag donuts in the main panel
2. Matching donut lenders onto information in the Avery file to assist the donut-auditing process
   1. Load our wide-format HMDA Crosswalk, and keep only the banks tagged with the “donut” variable from the step above
   2. Load the pre-2017 Avery file, and keep the following variables: RSSDyy, APPLyy, ORIGyy, ORIGDyy, ASSETLyy, ASSETyy
   3. Merge the Avery file data onto the donut-subset of the HMDA crosswalk
   4. Repeat steps ii. and iii. with the 2018 and 2019 versions of the Avery file
3. Use information from the Avery file to identify donut banks that are actually switching RSSD codes in certain years, and not non-reporting in those years
   1. Drop banks that never originate 100 mortgages in a year from the sample of banks to audit. This is consistent with HMDA rules indicating that a bank is required to report to HMDA if it originates at least 100 closed-end mortgage loans or 500 open-end lines of credit. If the bank never originates 100 closed-end mortgage loans, it is plausible that the bank truly did not file a HMDA report when it drops out of the HMDA dataset
   2. See Appendix B on the “donut hole filling” process. Out of 258 rows corresponding to donut banks, I find 6 rows where it is likely that a bank is actually filing under a different RSSD in the “non-reporting” years.

*At the end of this process, I reload the wide-format crosswalk. I identify 6 rows that have a reporting donut, which I believe actually report under a different RSSD code during years it appears the bank is not reporting. Each of these 6 rows is matched to another row by assigning the same masterid to each.*

5. Adjustments – adding banks in the loan-level data but not the lender panels, working with the 2020 and 2021 Avery files

1. Drop extraneous variables
2. Add in pre-2017 lenders that appear in the loan-level data, but not the lender panels
   1. In each of the 2013 and 2014 loan-level datasets, there are observations tagged with a HMDA lender ID that does not appear in the HMDA lender panels. I ultimately add these HMDA IDs in to our crosswalk - see Appendix B for my procedures to confirm that these HMDA IDs are assigned to the correct masterid.
3. Add in post-2017 lenders that appear in the loan-level data, but not the lender panels.

There are 138 LEI codes that appear in the post-2017 loan-level HMDA data but not the HMDA lender panels. Here, I attempt to match these LEIs onto RSSDs.

* 1. Load list of LEIs that do not appear in the lender panels (file was created from a previous attempt to merge the crosswalk onto the loan-level data, I have now saved it as a supplementary file in this replication package).
  2. Using LEI as a merge key, merge this set of banks onto:
     1. The NIC “Active” dataset
     2. The NIC “Closed” dataset
     3. The 2018 Avery file
     4. The 2019 Avery file
  3. Save a tempfile, and restrict to only the LEIs that have not matched to an RSSD yet
     1. Using LEI as a merge key, merge our set of unidentified LEIs onto LEIs in our HMDA ID Crosswalk. This handles cases where we have already matched an unidentified LEI to a masterid in one year, and through an error that same LEI is not included in the reporter panel in a different year
     2. Merge this subset back onto the full list of 138 LEIs
  4. For the LEIs that matched onto an RSSD in at least one of the files, check that none of the matches for a given LEI conflict with one another
  5. For the LEIs that found an RSSD, use RSSD to generate the masterid variable
  6. For the LEIs that did not find an RSSD, use LEI to generate the masterid variable (as above)
  7. Merge these banks onto the main panel, using masterid as a merge key

When the bulk of this code was written, the 2020 and 2021 Avery files were not yet published. What follows is a procedure to check whether these versions of the Avery file either 1) provide RSSDs for lenders for which we do not already have an RSSD associated (i.e. are identified with only an LEI or a metaid), or 2) contain RSSD information that conflicts with our panel

1. Work with the 2020 Avery file
   1. Merge 2020 Avery file onto our panel by HMDA ID, using concatid2020 as a merge key. In each row, store the RSSD we have recorded and the RSSD the Avery file has recorded in different variables
   2. Tag banks where the Avery file’s RSSD is non-zero/non-missing, and disagrees with our recorded RSSD (35 observations)
   3. For these 35 observations, use LEI to generate masterid, and perform an update merge back onto the main panel using masterid as the merge key
      1. 30 observations result in \_merge == 4. This means that of the 35 banks where we disagreed with the Avery file about what the RSSD for a given row should be, 30 of those were banks that we could only identify using LEI (that is, these were banks where we couldn’t find an RSSD). Now, we have RSSD information for those banks – we can “upgrade” these masterid’s to be RSSD-based.
      2. 5 observations result in \_merge == 2. This means that the remaining 5 banks are cases where we disagreed with the Avery file about what the RSSD for a given row should be, and we already were using RSSD to identify these banks. See appendix B – for 2 of these 5 banks, we’ve already manually recoded masterid (and in fact, the masterid we assigned agrees with the RSSD in the Avery file). For the remaining 3, I manually confirm that the RSSD we use to generate masterid agrees with the RSSD in the original HMDA lender panel. We leave these RSSDs unchanged and disagree with the Avery file.
   4. Save a tempfile of the banks we’ve upgraded to RSSD-based masterid (from step iii.1 immediately above) – we’ll merge these on shortly
2. Repeat the 2020 Avery file procedure with the 2021 Avery file
   1. After the merge described in d.iii immediately above, we have 45 \_merge == 4 observations and 29 \_merge == 2 observations. Once again, the \_merge == 4’s are banks we can now upgrade with new RSSD information. The \_merge == 2 observations I merge back onto the 2021 lender panel, and confirm that the RSSD we use to generate masterid agrees with the RSSD in the original HMDA lender panel. We leave these RSSDs unchanged and disagree with the Avery file.
3. Add our new information from the 2020 and 2021 Avery files onto our panel.
   1. In some cases, this new information means we are able to report a concatid2020/concatid2021 for an RSSD-identified bank we already had in the panel. In these cases, that same concatid2020/concatid2021 is already in our panel, but because it didn’t have RSSD information the first time, it existed as a stray row that was only identified by LEI. We “fill the hole” (populate the missing concatid for the RSSD-identified row), and delete the “stray” row to ensure we don’t create a duplicate
   2. In other cases, we are simply adding an RSSD to a bank that was previously only LEI-identified. There are no “strays” in these cases. So we just merge on the new RSSD information we’ve found, and change masterid to be equivalent to the RSSD.

6. Final cleanup – quality check, adding recoding flag, confirming bank names, variable management

1. Confirm that the concatid variables are unique within each year, using isid to check all observations where concatid is populated for a given year. This is a basic check, no HMDA identifier code should be associated with two different lenders in a single year.
2. Create a recoding flag variable. This is a binary variable which equals “1” in observations where I've changed the relationship between a HMDA ID/LEI and an RSSD according to personal judgement (as in step 2.a.i) or manually changed a row’s masterid (as in step 4.d). The complete list of masterids where this flag is appropriate can be found at the top of Appendix C, and this step is marked with the comment “6.b” in the code.
3. Re-attach bank names. In extremely rare cases, I have found that a given name[yyyy] variable is not the name associated with the corresponding concatid[yyyy]. Out of an abundance of caution, we go back and merge our panel with each year’s lender panel using concatid[yyyy] as a merge key in each year. This ensures that our panel contains the correct name associated with a bank’s HMDA identifier in the lender panel in each year.
4. Drop remaining extraneous variables, assign variable labels, order and sort dataset.

END OF PROCEDURE

## 8. Sources

## Appendix A: Codebook

**masterid**

*masterid* is the primary identifier in the dataset – this is the single code that we argue defines an economically distinct lender. Each *masterid* is linked to the time series of codes used to identify a given lender in the HMDA lender panels and loan-level data in each year (these codes are contained in the *concatid*[yyyy] variables, see below).

See section 2ii for a description of the process by which we construct *masterid*.

Though most rows are unique in terms of *masterid*, there are 33 pairs of rows with duplicated values of *masterid* – see section 5 of this documentation for details.

**rssd**

RSSD codes are the primary variable we use to construct *masterid*, though RSSD codes are not available for all banks.

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.”[[8]](#footnote-8) 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.[[9]](#footnote-9)

See sections 2 and 4 for further discussion.

In the rare cases that there are two rows corresponding to a single *masterid*, those two rows usually contain distinct RSSDs which are associated with their own time-series of *concatid*[yyyy] variables. The duplicated *masterid* occurs because we judge that the two RSSDs correspond to a single lender.

**metaid**

In cases when neither RSSD nor LEI are available to identify a lender, we construct an ad hoc identifier called *metaid* which serves as the basis for the *masterid* variable.

**concatid[yyyy] (where yyyy=2010-2021)**

The *concatid*[yyyy] series of variables contain the code used to identify a given lender in HMDA in each year. In other words, *masterid* is the code that distinguishes an individual lender, and the *concatid*[yyyy] for a given year contains the code to identify that lender in that year’s HMDA data. To link *masterid*-identified banks to a given year’s loan-level data, the user must use the *concatid* variable corresponding to that year as a merge key between the HMDAHarmonizer panel and the loan-level data.

In the pre-2018 loan-level data, lenders are identified in the HMDA data with the combination of two variables: agency code and respondent ID (see Section 2). For years 2010-2017, the *concatid*[yyyy] variables are concatenated strings of these two variables, and must be processed before useable. Agency code is always stored as the first character in the string, and the remaining characters contain the respondent ID. Thus, the user must split each *concatid* variable into an agency code and a respondent ID and use the combination of these two variables in the merge. See example code in Appendix B.

For years 2018-2021, banks are identified in the HMDA data using LEI codes (see Section 2). the concatid[*yyyy*] variables in these years are the exact LEI code, no processing is required to use these variables to merge *masterid* onto loan-level data.

**name[yyyy] (where yyyy=2010-2021)**

The *name*[yyyy] series of variables contain the name used to represent a lender in HMDA in a given year.

*masterid* is the harmonized code identifying a bank over time, and it is linked to a series of *concatid*[yyyy] variables. Each *concatid*[yyyy] corresponds to a *name*[yyyy] variable in the same year. We pull both these pairs from each year’s HMDA lender panel. Thus, the series of *name*[yyyy] variables shows how the lender is literally titled in the HMDA lender panels in each year.

**recoding\_flag**

Binary variable to indicate whether a bank (as identified by *masterid*) ever experiences some type of recoding in a manual audit. Specifically, this variable will equal “1” in cases when:

* The relationship between an RSSD and a HMDA ID (either pre-2018 HMDA ID or LEI) is manually recoded
* We manually recoded a bank’s masterid (as in our procedure to identify RSSD switchers)

The steps when these recodings occur are all explicitly marked in the step-by-step description and section 7. The code includes a comment giving the number of the step when these recodings occur (for example, users can search “2.c.iii.1” in the code, and will find the appropriate spot). The basis behind each recoding is described in Appendix B.

The full set of masterids for which this variable equals 1 is:

* 3343717 4320395 568443 (recoded in step 2.a.i)
* 3075401 (step 2.b.ii.5)
* 3874118 4327965 (step 2.b.ii.8)
* 102379 (step 2.c.iii.1)
* 867856 3383665 (step 3.e.ii.2)
* 5019678 3876710 980951 649397 64897 3310456 3878750 111205 251978 61122 200378 2760232 754929 3195242 713926 968557 977951 (step 3.f)
* 4455073 2351078 3844492 3882560 3950469 3876390

2383060 3715220 3720532 (step 4.a.v-4.a.vi)

* 112837 959304 3913633 1216826 2860459 672984 (step 4.d)

## 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)
8. [About - National Information Center (ffiec.gov)](https://www.ffiec.gov/npw/Home/About) [↑](#footnote-ref-8)
9. NIC Data Dictionary, p. 21 [↑](#footnote-ref-9)