A New Loan-Level Data Set: HMDA$^+$

Chao Liu*

August 21, 2021

There are many mortgage data sets containing important information about some aspects of a mortgage or borrower in the U.S., but none of them offers a full set of loan and borrower characteristics. For example, Home Mortgage Disclosure Act (HMDA) data cover a large sample and contain information about borrowers' demographic characteristics, but the HMDA data lack underwriting and performance variables (e.g., FICO scores, interest rates, and delinquency rates). The National Survey of Mortgage Originations (NSMO) data offer rich and unique information about borrowers' mortgage market experiences, but this data set does not provide location or lender information. Fannie Mae and Freddie Mac Single-Family Loan Level Data (GSE data) contain mortgage contract variables, but they have little information on borrowers' race and income.

The key problem in the study of U.S. mortgage market is that there is no unique loan identification number to connect existing mortgage data sets. To overcome this data challenge, I construct a novel loan-level data set, HMDA$^+$, which contains the most detailed loan-level information of borrowers, lenders, mortgages, and properties. Compared to existing efforts (Saadi, 2020; Bartlett et al., 2022), my HMDA$^+$ data set has three advantages. First, I only use publicly available mortgage data to assemble the final sample. Second, I exploit lenders' names in the matching step to achieve a higher matching rate. Finally, analysis weights are developed to represent the population of originated mortgages in HMDA data.

The benchmark data source is the HMDA (2015-19) data set, as it covers the near universe of originated mortgages in the U.S. After merging with the Avery files, HMDA data can provide both borrowers' demographic characteristics and lenders'

*Liu: Kellogg School of Management, Northwestern University. Email: chao.liu1@kellogg.northwestern.edu.

identities.[1] Therefore, what is missing is underwriting and performance information. For mortgages sold to GSEs, I will retrieve these variables from the Fannie Mae and Freddie Mac data sets. For FHA loans, I will use the Ginnie Mae data disclosure, which is similar to the GSE data set in both format and content. To match the HMDA data set with other mortgage data sets, I mainly rely on property locations, origination time, and some basic mortgage characteristics, such as loan size and loan purpose. The more granular the location information is, the better the matching results are.

A Byproduct: FHA Snapshot$^+$

The most granular geographical level in Ginnie Mae data is at the state level, which is too coarse for precise matching with the HMDA data. Therefore, I first match FHA Snapshot data with Ginnie Mae data to supplement detailed location variables.

I restrict my sample to fixed rate, purchase or refinance FHA loans that were originated after 2015 in the Ginnie Mae data set. I merge the sample with FHA Snapshot data based on a set of overlapping variables: state, loan purpose, origination year and month, lender, loan size, and interest rate. Specifically, I allow a 2% difference in loan amount and a 3-month difference in origination month in the first round of fuzzy matching. For each one-to-many match, I keep the match with the smallest size and time difference. To ensure the highest-quality match, I exclude all matches with duplicate observations. The origination time in FHA Snapshot is systematically later than that in Ginnie Mae data. Therefore, I select unmatched mortgages that were originated in the fourth quarter after the first round of matching in Ginnie Mae data, and then I merge them with next year's FHA Snapshot data. I further drop mortgages with missing FICO scores or LTV ratios, and only keep mortgages with a loan term of 120 months, 180 months, 240 months, or 360 months. Following the above steps, the final data set contains 1,783,367 FHA loans originated from 2015 to 2019. Figure 1 shows that the FHA Snapshot$^+$ data set covers about 32% of all FHA loans and 65% of FHA loans sold to Ginnie Mae in HMDA data.

To make this sample representative, I use the reciprocal of the likelihood of being sampled from HMDA data as the analysis weight. This assumes that each mortgage in the FHA Snapshot$^+$ data set is randomly sampled from the corresponding stratum in the HMDA data of FHA loans. I separate all FHA loans in HMDA into different

[1] You can download the Avery files from https://sites.google.com/site/neilbhutta/data.

strata based on property county, loan size, loan purpose, and origination year. In other words, the analysis weight indicates how many FHA loans of a certain type in the mortgage market are represented by a given loan in the FHA Snapshot$^+$ sample. Table 1 shows that the mortgage characteristics in the FHA Snapshot$^+$ are very similar as those in Ginnie Mae data.

B Final Sample: HMDA$^+$

The final data set, HMDA$^+$, uses three data sources: HMDA, GSE data, and FHA Snapshot$^+$. To construct this data set, I first merge Fannie Mae data with HMDA data. I pick one to four-family mortgages sold to Fannie Mae in HMDA data for match. I merge these loans with Fannie Mae data based on loan amount, MSA, 3-digit ZIP code, state, lender, loan purpose, occupancy status, and the number of borrowers. As HMDA started to disclose mortgage rate after 2018, I also use this variable as a match key. To accommodate for rounding differences, I allow a 2% difference in loan amount. For each one-to-many match, I keep the matched pair with the smallest loan size difference. To ensure the highest-quality match, I further exclude all matches with duplicate observations. Because only large lenders are identified in Fannie Mae data, a large amount of Fannie Mae loans will not be matched after the above steps. Therefore, I match the leftover mortgages in both data sets without lender identification. Similarly, I repeat the above steps for mortgages sold to Freddie Mac in HMDA. The final data set contains 8,154,065 GSE loans originated from 2015 to 2019, which covers about 59% of GSE loans in HMDA data. Figure 2 shows that the matching rate increases when interest rate is used for matching.

The second step is to merge FHA Snapshot$^+$ with HMDA. I select one to four-family FHA loans in HMDA for match. I merge two data sets based on county, 5-digit ZIP code, loan purpose, origination year, loan amount, and lender. Interest rate is used as an additional match key after 2018. To accommodate for rounding differences between the two data sets, I allow loan amount to differ by 2% and interest rate to differ by 2 basis point. For each one-to-many match, I keep the match with the smallest size and interest rate difference. To ensure the highest-quality match, I further exclude all matches with duplicate observations. Using this approach, the final data set contains 1,376,241 FHA loans originated from 2015 to 2019, which covers roughly 25% of FHA loans in HMDA data during the same period. The matching rate is lower than that for GSE loans because the baseline data set, FHA Snapshot$^+$, only covers about 32% of all FHA loans. Overall, the HMDA$^+$ data set includes nearly half of GSE and FHA loans in the 2015-19 HMDA data.

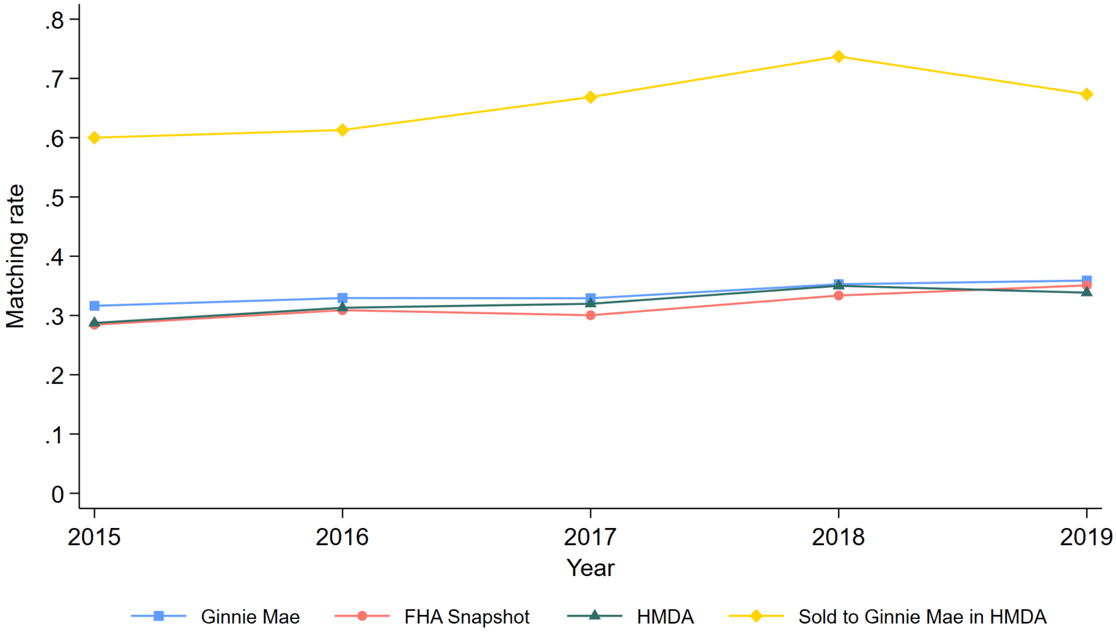
To make the HMDA$^+$ sample representative, I add an analysis weight for each observation, using the reciprocal of the likelihood of being sampled from the HMDA data. The sampling procedure assumes a random sampling within the corresponding stratum. I separate all originated loans in HMDA data into different strata based on property county, loan type, loan size, loan purpose, and origination year. As shown in Table 2, the HMDA$^+$ data set represents the whole mortgage market quite well.

References

Bartlett, Robert, Adair Morse, Richard Stanton and Nancy Wallace. 2022. "Consumer-Lending Discrimination in the FinTech Era." *Journal of Financial Economics* 143(1):30–56.

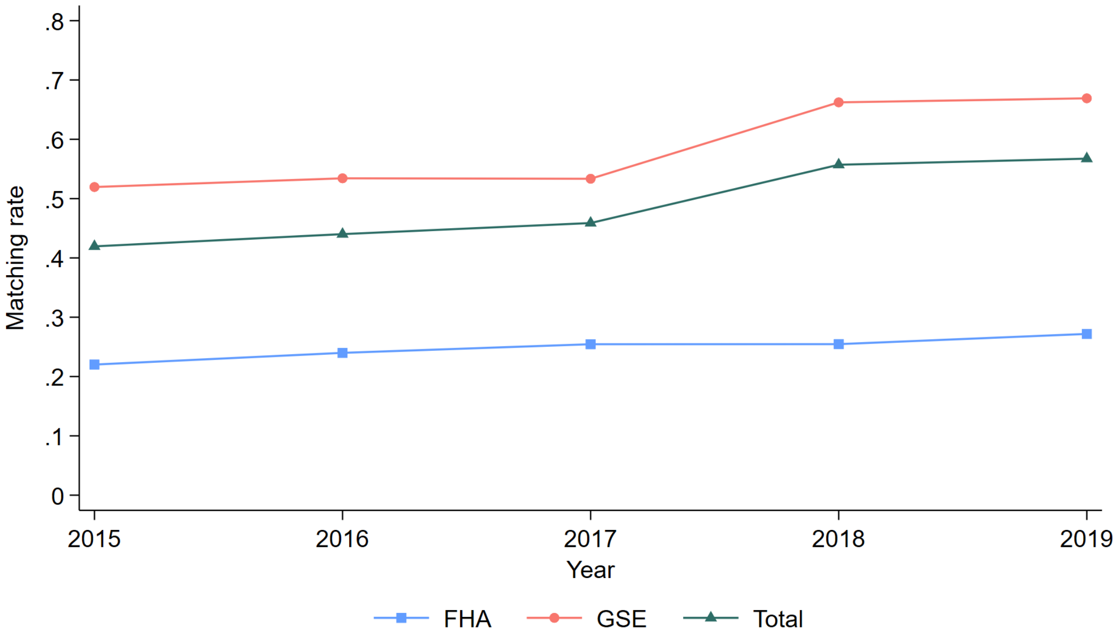
Saadi, Vahid. 2020. "Role of the Community Reinvestment Act in Mortgage Supply and the U.S. Housing Boom." *The Review of Financial Studies* 33(11):5288–5332.

Figure 1. FHA Snapshot$^+$ Matching Rate



Notes: This figure plots the matching rates between the FHA Snapshot$^+$ dataset and Ginnie Mae data, FHA Snapshot data, FHA loans in HMDA, and FHA loans sold to Ginnie Mae in HMDA.

Figure 2. HMDA^{+} Matching Rate



Notes: This figure plots the matching rates between the HMDA^{+} dataset and Ginnie Mae data, FHA Snapshot data, FHA loans in HMDA, and FHA loans sold to Ginnie Mae in HMDA.

Table 1. Summary Statistics of the FHA Snapshot$^+$ Sample

Sample	FHA Snapshot$^+$ (1)	Ginnie Mae (2)
Purchase	0.753	0.745
	(0.431)	(0.436)
Interest rate	4.159	4.176
	(0.586)	(0.590)
Loan amount($1K)	209.186	205.926
	(104.120)	(103.770)
LTV	92.956	93.100
	(9.370)	(9.150)
DTI	42.016	41.962
	(9.417)	(9.288)
FICO scores	671.276	676.082
	(50.038)	(49.031)
Observations	1,783,367	5,298,341

Notes: This table reports descriptive statistics of the FHA Snapshot$^+$ dataset (column 1) and the Ginnie Mae dataset (column 2). All table entries represent sample means and standard deviations in parentheses. Summary statistics in column 1 are weighted by the the reciprocal of the likelihood of being sampled from the HMDA data of FHA loans.

Table 2. Summary Statistics of the HMDA$^+$ Sample

Sample	HMDA$^+$ (1)	HMDA (2)
GSE loans	0.726	0.714
	(0.446)	(0.452)
Purchase loans	0.581	0.581
	(0.493)	(0.493)
Owner occupied	0.905	0.914
	(0.293)	(0.281)
Loan amount	230.351	228.073
	(118.585)	(135.285)
Female	0.339	0.340
	(0.473)	(0.474)
Black	0.073	0.072
	(0.261)	(0.259)
Asian	0.070	0.067
	(0.255)	(0.249)
Hispanic	0.122	0.126
	(0.328)	(0.331)
Income	101.351	100.722
	(85.129)	(444.782)
Observations	9,530,306	19,449,814

Notes: This table reports descriptive statistics of the HMDA$^+$ dataset (column 1) and the HMDA dataset (column 2). All table entries represent sample means and standard deviations in parentheses. Summary statistics in column 1 are weighted by the the reciprocal of the likelihood of being sampled from HMDA data.