NMDB
National Mortgage Database

cfpb Consumer Financial Protection Bureau FHFA Federal Housing Finance Agency

Technical Report Series

National Mortgage Database

Technical Report 15-01

August 17, 2015

1.0   Introduction

The National Mortgage Database project is a multi-year project being jointly undertaken by the Federal Housing Finance Agency (FHFA) and the Consumer Financial Protection Bureau (CFPB). The project is designed to provide a rich source of information about the U.S. mortgage market based on a five percent sample of residential mortgages. It has two primary components: (1) the National Mortgage Database (NMDB) and (2) the quarterly National Survey of Mortgage Borrowers (NSMB).

The NMDB will enable FHFA to meet the statutory requirements of section 1324(c) of the Federal Housing Enterprises Financial Safety and Soundness Act of 1992, as amended by the Housing and Economic Recovery Act of 2008, to conduct a monthly mortgage market survey. Specifically, FHFA must, through a survey of the mortgage market, collect data on the characteristics of individual mortgages, including those eligible for purchase by Fannie Mae and Freddie Mac and those that are not, and including subprime and nontraditional mortgages. In addition, FHFA must collect information on the creditworthiness of borrowers, including a determination of whether subprime and nontraditional borrowers would have qualified for prime lending.[1]

For CFPB, the NMDB project will support policymaking and research efforts and help identify and understand emerging mortgage and housing market trends. The CFPB expects to use the NMDB, among other purposes, in support of the market monitoring called for by the Dodd-Frank Wall Street Reform and Consumer Protection Act, including understanding how mortgage debt affects consumers.

FHFA and CFPB considered existing databases but determined that none sufficiently support the above objectives.[2] The NMDB, when fully complete, will be a de-identified loan-level database of closed-end first-lien residential mortgages. It will: (1) be representative of the market as a whole; (2) contain detailed, loan-level information on the terms and performance of mortgages, as well as characteristics of the associated borrowers and properties; (3) be continually updated; (4) have an historical component dating back before the financial crisis of 2008; and (5) provide a sampling frame for the NSMB (see NMDB Technical Report 15-02).

The core data in the NMDB are drawn from a random 1-in-20 sample of all closed-end first-lien mortgage files outstanding at any time between January 1998 and June 2012 in the files of Experian, one of the three national credit repositories.[3] The use of a sampling frame substantially reduces the privacy risk associated with any data collection. By contrast, a universal registry can present challenges for privacy since it is known that a particular loan must be in the dataset. However, for a 1-in-20 sample, the odds are 95 out of 100 that a particular loan is not in in the

---

[1] FHFA interprets the NMDB project as a whole, including the NSMB, as the "survey" required by the Safety and Soundness Act. The statutory requirement is for a monthly survey. Core inputs to the NMDB, such as a regular refresh of credit-bureau data, occur monthly, though the NSMB does not.

[2] Please see the Appendix for a discussion of existing sources and their limitations.

[3] Experian was chosen through a competitive procurement process to assist in creating the NMDB.

database. In addition, the sample used is large enough to support almost all types of statistically valid analyses but small enough to manage logistically, thus dramatically reducing both contract and personnel costs.

A random 1-in-20 sample of mortgages newly reported to Experian is added each quarter. Mortgages are followed in the NMDB database until they terminate through prepayment (including refinancing), foreclosure, or maturity. Information from credit repository files on each borrower associated with the mortgages in the NMDB sample is collected from at least one year prior to origination to one year after termination of the mortgage. The information on borrowers and loans available to the FHFA, CFPB, or any other authorized user of the NMDB data is de-identified and does not include any direct identifying information such as borrower name, address, or Social Security number.

This technical report is designed to provide users of the NMDB data with background on the development of the database, as well as an assessment of the quality of its data. The remaining sections of this report discuss the development of the contract with Experian, outline the process of selecting the initial historical sample, describe how the initial sample data were processed, discuss how the data are being updated, and how administrative data are being merged into the NMDB. The final section then evaluates the NMDB sample frame.

2.0   The Experian Contract

By interagency agreement between FHFA and CFPB, FHFA leads the production of the NMDB. Following a competitive procurement process, a five year contract for the core data of the NMDB was signed between FHFA and Experian in September 2012. Simultaneously, FHFA and CFPB signed an interagency agreement that codified the cost-sharing (shared equally) and administrative arrangement.

The Experian contract has several key elements designed to ensure compliance with the Fair Credit Reporting Act (FCRA) and to protect the privacy of both borrowers and lenders.[4] First, while Experian will be using name, address and Social Security number for matching purposes only, this information will not be transmitted to FHFA or CFPB when constructing the NMDB. Second, any user of the database must sign a "terms of use agreement" that states that they will not attempt to learn the identity of any borrower.[5] Third, all access to the NMDB must be through a server at FHFA or CFPB and strictly controlled. Fourth, the NMDB – which is

---

[4] The Fair Credit Reporting Act (FCRA), Public Law No. 91-508, was enacted in 1970, and substantially amended since, to promote accuracy, fairness, and the privacy of personal information assembled by credit reporting agencies (CRAs). The Act's primary protection requires that CRAs follow "reasonable procedures" to protect the confidentiality, accuracy, and relevance of credit information. To do so, the FCRA establishes a framework of requirements for credit report information that include rights of data quality (right to access and correct), data security, use limitations, requirements for data destruction, notice, user participation (consent), and accountability.

[5] Though FHFA and CFPB have not yet established policies of access or determined who may attempt to obtain access, the contract allows access to the NMDB to be extended to employees of other federal agencies, the Federal Reserve System, Fannie Mae, Freddie Mac, and Federal Home Loan Banks, provided the employee has signed the terms of use agreement.

designed to describe the market as a whole – cannot be used for enforcement against any specific servicer or lender.

3.0   Selecting the Initial Sample

The credit repository core of the NMDB is being developed in two phases: (1) an initial 1-in-20 random sample of closed-end first-lien mortgages active at any time from January 1998 to June 2012 (January 1998 was the earliest available date given Experian's archive policies); and (2) quarterly updates that add a 1-in-20 random sample of mortgages newly reported to Experian and updated information on existing loans still active in the database.

One of the virtues of the credit repository sampling frame is that the repositories maintain records in a credit report not only of mortgages (and other credit obligations) that are currently active, but also of those that are closed. However, because of the FCRA, records with derogatory information are purged from the current credit report after seven years from their point of first continual delinquency, and Experian's policies dictate a purge of all closed accounts 10 years after their closing.

However, since Experian retains archives of their data for 10 years or longer, data on mortgages that have been purged from Experian's current files can be recovered. These archives, which are not used for credit granting decisions, contain snapshots of each credit record as it existed at the close of business on a given day of each month, except that personal information, such as name, address, and Social Security number, is suppressed.

The bulk of the initial sample for the NMDB was drawn from the June 2012 archive. This was supplemented by samples from the December 2005 and July 2001 archives that captured loans that may have been purged from the current files by June 2012.

Trade lines, which are records that contain information about specific loans or debt obligations that are reported by loan servicers, account for most of the information contained in credit records. Loan servicers typically update trade line information on a monthly basis using a standardized format agreed upon by the servicers and the credit repositories (Metro 2® format). The updates include information on the opening date of the loan, the current and original loan balance, the type of servicer, loan term and type, payment amount, and loan repayment performance.

However, the format agreed upon by loan servicers and the credit repositories does not perfectly identify closed-end first-lien mortgages. Recognizing that some second liens would be sampled and have to be removed later, trade lines falling under the following categories were deemed eligible for the NMDB:

> (1) any trade line with a Metro 2 "Enhanced Account Type Code" of: 08 (Real estate loan, specific type unknown), 19 (FHA real estate mortgage), 2C (FMHA real estate mortgage) , 25 (VA real estate mortgage), 26 (Conventional real estate mortgage), 27 (Real estate mortgage, with or without collateral, usually second mortgage), 85 (Bi-monthly mortgage payment), 87 (Semi-monthly mortgage payment), 5A (Real estate –

4

junior liens and non-purchase money first), 17 (Manufactured home loan) , and 05 (FHA home-improvement loan); or

(2) trade lines reported by servicers with "Kind of Business Codes" of: FB (Mortgage Brokers), FM (Mortgage Companies), FR (Mortgage Reporters), RE (Real Estate Sales and Rentals), BM (Bank-mortgage only), FL (Savings and loan – mortgage department) and Metro 2 "Enhanced Account Type Codes" of: 02 (Secured loan), 04 (Home improvement loan), 66 (Government- secured guaranteed loan), 7B (Agriculture), 9A (Secured home improvement) or a "Secondary Agency Code" of: 01 (Fannie Mae) or 02 (Freddie Mac).

Trade lines in the June 2012 archive that met either of the above criteria were included in the population from which the initial NMDB 1-in-20 random sample of mortgages was drawn. Any open-ended or revolving loans otherwise meeting one of the criteria were excluded from the sampling universe. No other restrictions were imposed.

The first supplemental sample was a 1-in-20 random sample of trade lines drawn from the December 2005 archive that met the criteria for the June 2012 archive, had information reported for some period in the past 7 years (indicated by an "Account Balance Date" of January 1998 or later), and were opened in September 2005 or earlier. In order to exclude loans from the 2005 sample that should be present in the June 2012 archive, loans were excluded if they were last reported after July 2002 with a reported account status of "current."

The second supplemental sample, drawn from the July 2001 archive, was a random 1-in-20 sample of trade lines that met the criteria used for the June 2012 archive and that had "Account Balance Dates" of January 1998 or later and "Account Open Dates" of April 1999 or earlier. Any trade line with an "Enhanced Status Code" of "current" was excluded from the sample. Again, these additional conditions were designed to exclude from the 2001 sample all trade lines that should be present in the 2005 archives.

4.0   Processing the Initial Sample

For each archival pull, all available individual depersonalized credit records, including trade lines, inquiries, and public records (collectively, TIPs) associated with all borrowers accompanying any initial sample trade line were provided regardless of the archive from which it was sampled. The data provided by Experian are de-identified and contain no directly identifying personal information such as name, address, or Social Security number. The credit records were tagged with de-identified borrower numbers (DINs) and servicer and loan numbers (both in encrypted form).[6] These could be used (imperfectly) to link TIP files to other account-level files both within an archive and over time.

---

[6] The encrypted servicer identification and loan numbers are used only by the NMDB development team primarily to update the database each quarter. They are not available to dataset users even in encrypted form. This is done to ensure compliance with the contract restriction that the database not be used for enforcement against servicers. The borrower DINs are unique to the NMDB and are randomized.   Experian, however, maintains the mapping between the borrower identification numbers used in their system and the DINs supplied to the NMDB team so that records in the NMDB associated with the same DIN will be associated with the same borrower ID in the Experian records.

One major problem encountered with the NMDB sample frame is that a single mortgage can be associated with multiple trade lines. This can arise when the servicing of the loan is sold or transferred and the trade line reported by the original servicer is not properly linked to the trade line reported by the new servicer. In such cases, borrowers may appear to have multiple mortgages, when, in fact, they have only one. Because of these duplicates, randomly sampling trade lines will result in mortgages with multiple records being over represented in the data. To correct for this, a processing methodology was developed to identify and combine multiple records that contain information about the same mortgage into one record.

The first step in the process of eliminating duplicate mortgage records ("de-duping") was to find multiple trade lines for the same mortgage in the same archive. From these duplicates, sample loans were removed when the selected trade line was not the one with the latest "Account Balance Date" (this corrects for the problem of having mortgages associated with multiple trade lines over-represented in the sample). The second step was de-duping across archives. The June 2012, December 2005, and July 2001 samples were treated as sequential NMDB sample frames (in that order) whereby mortgages selected from a NMDB sample frame later in the order (*e.g.*, July 2001) that can be found in a NMDB sample frame earlier in the order (June 2012 or December 2005) would be removed from the sample (again, this corrects for the fact that such mortgages are over-sampled in the raw frame).

The de-duping process also dealt with the problem of ambiguous lien status for the "Enhanced Account Type Codes" of 08 (Real estate, specific type unknown), 27 (Real estate mortgage, with or without collateral, usually second mortgage), and 5A (real estate – junior liens and non-purchase money first). Sample trade lines associated with these codes were removed from the sample when they subsequently could be linked with trade lines that were unambiguously second liens.

Once the initial samples were de-duped, it was necessary to link archival records over time to create a composite picture of the performance of each sample loan. Semi-annual archives were drawn for the period December 2001 to December 2011 for borrowers associated with the initial sample loans. Data from these archives were patched together to create a temporal picture of each loan. One issue that needed to be dealt with is that DINs for a given borrower can change over time. There are times when a loan is first reported to the credit repositories and cannot be connected with existing credit records for the borrower(s). This can happen because lenders make errors in reporting names and addresses or because of changes to a borrower's addresses or names. In this instance Experian treats the loan as associated with a new borrower. In most of these instances the records are ultimately reconciled with the correct existing borrower and a "DIN-merge" occurs. However, historical archives are stored with the DINs at the time of the archive. Thus, to properly connect borrowers (and mortgages) over time, it was necessary for Experian to provide a DIN-merge transformation table to map historical to current DINs.

As shown in Table 1, the de-duping process substantially reduced the size of the original NMDB sample. About 15 percent of the mortgage trade lines originally sampled from the June 2012 archive, more than a quarter of the selections from the 2005 archive, and almost three-quarters of the selections from the 2001 archive were dropped. The percentages were higher for the older archives since many of the loans selected from them were selected because they were not current

at the date of the archive and thus subject to the FCRA purge rules. However, many of these loans subsequently became current and could be found in later archives.

Table 1				
Archive Date	Sample Tradelines	Final Loans	Final Borrowers	Percentage Dropped
July 2001	302,398	86,797	133,127	71.3 %
Dec 2005	2,955,675	2,158,188	3,520,538	27.0
June 2012	9,225,304	7,794,176	12,169,729	15.5

5.0 Updating the Sample

Under the NMDB sample design going forward, credit records for borrowers associated with sampled mortgages are to be collected quarterly until one year after the mortgage is reported as closed.[7] As of June 2012, approximately 3 million loans from the initial sample were still active or had been closed less than a year. In addition, to keep the NMDB up-to-date, it is necessary to add a representative sample of the new mortgages reported to Experian each quarter to the database.

The initial update of the NMDB from the June 2013 archive covered a full year of newly-reported mortgages since June of 2012. Since that date, updates have taken place quarterly drawing from the last archive of the quarter (March, June, September or December). Each quarterly update follows the same pattern. A 1-in-20 random sample of closed-end first-lien mortgage trade lines is drawn. These loans, which are identified using the same criteria as was used for the June 2012 archive, are selected from among the loans that were newly reported to Experian since the date of the previous quarterly update archive. The new sample is de-duped using the same methodology as used for the initial sample. If multiple trade lines are identified for the mortgage and the selected mortgage is not the one with the latest "Account Balance Date" or the mortgage is deemed to be a second lien then it is dropped. In addition, checks are run to determine if the mortgage was already reported in an earlier archive period (perhaps as a different trade line). If so, the loan is dropped.

Existing sample loans are also updated each quarter. Prior to the update, the DIN-merge transformation table is updated to account for "newly merged" DINs. To ensure that lagged information for all DINs newly added to the dataset is collected, the year-old archive is drawn each quarter for all active DINs for which this archive had not previously been collected.

At present between 75,000 and 80,000 new loans are added to the NMDB each quarter (see Table 2). The number of mortgages added to the database is only about two-thirds of the raw trade lines originally selected for the update sample.

---

[7] A partial update is done monthly collecting only limited performance data for active sample mortgages. This allows the database to provide high-frequency information on mortgage delinquency rates.

		Table 2		
Archive Date	Sample Tradelines	Final Loans	Final Borrowers	Percentage Dropped
June 2013	648,224	499,466	775,732	22.9 %
Sept. 2013	240,001	132,336	201,641	44.9
Dec. 2013	174,404	110,326	163,897	36.7
Mar. 2014	111,928	54,564	80,962	51.3
June 2014	146,406	79,800	118,042	45.5
Sept. 2014	124,389	76,911	114,294	38.2
Dec. 2014	124,323	77,792	115,078	37.4
Mar. 2015	104,613	75,284	111,859	28.0
June 2015	129,737	93,822	139,886	27.7

6.0 Merging with other Data Sources

Although extensive, Experian's archive files do not contain information on a number of key mortgage features, such as the loan's purpose (home purchase or refinance), whether it had an adjustable or fixed rate, its securitization status, its origination channel (broker or retail lender), or whether it was for an owner-occupied property, vacation home or investor property. Moreover, Experian's archives contain no information on the property backing the mortgage, such as its location, purchase price, characteristics, or current value. Finally, key information on borrowers associated with the loan including income is also missing. Consequently, values of these key variables need to be inferred indirectly or acquired from other data sources if they are to be included in the NMDB.

The NMDB expects to obtain much of the missing information from matches to administrative file records. Predominantly the administrative files come from government-affiliated mortgage programs including Fannie Mae and Freddie Mac (the Enterprises), and tentatively, the Federal Housing Administration (FHA), and the U.S. Department of Veterans Affairs (VA). Collectively, loans associated with these programs comprise about three-quarters of the loans in the NMDB.

The most accurate means of merging information from outside sources into the NMDB would be to use information about the borrowers, such as their names, Social Security numbers, addresses, and dates of birth. Using such directly identifying information (DII), however, would heighten concerns about data security and borrower privacy. Consequently, FHFA contracted with an outside consultant to conduct a study of how such concerns might be mitigated. The third-party-blind matching process that FHFA used is consistent with the "best practices" and recommendations from that study.
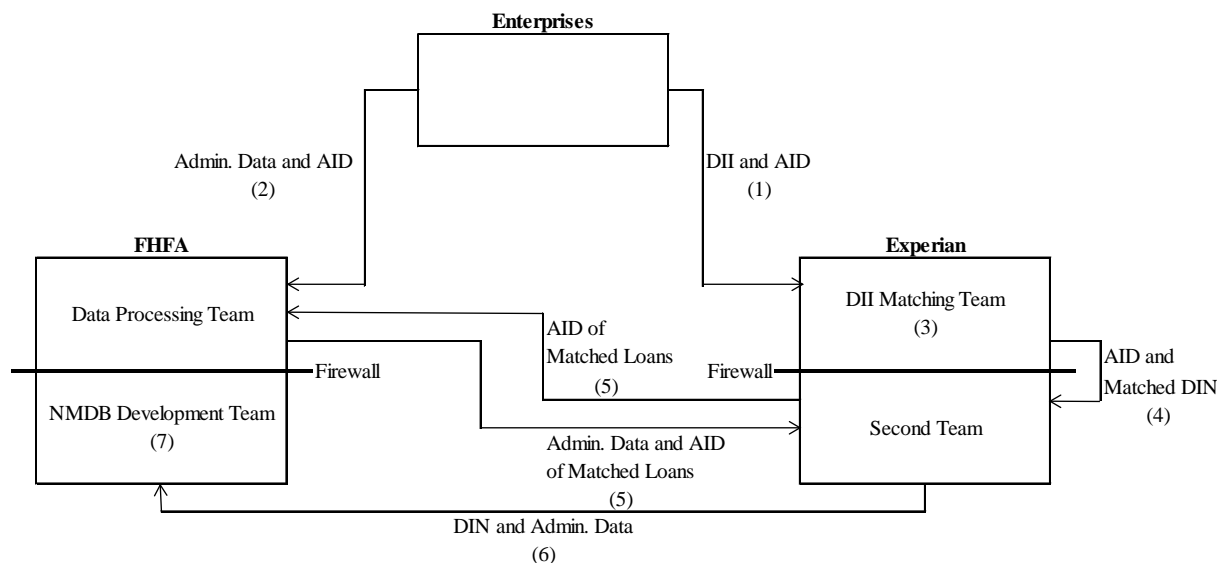
The third-party-blind matching process adheres to three guiding principles. First, neither FHFA nor the Enterprises can receive DII from Experian. Second, Experian cannot access both Enterprise administrative data and borrower DII in the same place. Third, FHFA must not be able to match loans in the NMDB records to the specific administrative records from the Enterprises.

In December 2014, a process was initiated to supplement the NMDB data with administrative data from Fannie Mae and Freddie Mac. The process for matching the data from the Enterprises followed seven steps:

(1) The Enterprises created a unique anonymized identifier (AID) for each loan. This identifier, along with the borrower-level DII associated with each loan (including name, address, Social Security number, and date of birth), was transmitted directly to Experian using a secure portal. FHFA did not receive this information. Other administrative data on these loans were not sent to Experian.

(2) The Enterprises sent the AID, along with administrative data for each loan, to an FHFA data processing unit that is separate from the NMDB development team. No borrower-level DII was included in the information sent to the FHFA data processing unit.

(3) Behind a secure firewall to protect FCRA-regulated data, Experian matched the DII it received from the Enterprises to the DII maintained in its own files on the borrowers in the NMDB to determine potential matches. When a potential match was identified, Experian compiled the DIN for each matched borrower.

(4) For all potential matches, Experian transferred the Enterprise-supplied AID and the matched NMDB borrower DINs to a separate unit within Experian that had no access to the credit repository data or any DII.

(5) The second Experian unit sent the list of matched AIDs to the data processing unit in FHFA that received the administrative data from the Enterprises in step (2). For each AID it received, this data processing unit sent back the associated administrative data that it received from the Enterprises.

(6) After receiving this information, the second Experian unit forwarded the administrative data they received from the data processing unit at FHFA, plus the matched borrower DIN that they received from the first Experian unit, to the NMDB development team at FHFA. The information sent to the NMDB development team included neither the Enterprise-created AID nor any DII.

(7) The NMDB team compared the characteristics of the loans associated with the DINs received from the second Experian unit to the administrative information on the loans. If the information from both sources was consistent, the match was confirmed. A list of confirmed matches was sent to Experian. Upon confirmation, Experian stored the property address supplied as part of the DII file from the Enterprises but otherwise permanently destroyed all DII used in the match.

The figure below illustrates the third-party-blind matching process.

NMDB Administrative Data Merging Process

Enterprises

Admin. Data and AID (2)

DII and AID (1)

FHFA

Data Processing Team

NMDB Development Team (7)

Firewall

AID of Matched Loans (5)

Firewall

Admin. Data and AID of Matched Loans (5)

DIN and Admin. Data (6)

Experian

DII Matching Team (3)

AID and Matched DIN (4)

Second Team

As of this writing, results of the Enterprise administrative file matching are still being processed. Negotiations are also underway to use similar methods to match FHA and VA loans with the NMDB. Contracts have also been signed to merge property record information into the NMDB, using similar third-party blind matching techniques. Data from servicing and private-label databases will also be matched which should provide missing data elements for most of the non-government-affiliated loans in the NMDB.[8]

It is anticipated that additional matching will be conducted to enhance the NMDB with information from the Home Mortgage Disclosure Act (HMDA) data, private mortgage insurance companies, the Rural Housing Service, and the Federal Home Loan Banks.[9] These matches will likely not involve DII and those will have to reply on less accurate techniques.

Ultimately, the NMDB will combine data from all of these sources into a common file with one record per sampled loan. The record will contain variables reflecting all the static characteristics

---

[8] To facilitate the property matching, the entire property database of one of the two largest U.S. property data vendors has been placed behind the secure firewall at Experian. This allows information on borrower name and address to be used in the matching process. Again, any DII used in the match will be discarded once the matching process is completed.

[9] Such merges will use information common to both datasets to perform a match but not DII. Most of the matches contemplated for the NMDB will rely on the original loan balance, the opening date of the mortgage and the general location of the property (census tract, ZIP Code or state/county). Unfortunately, mortgage servicers report the billing address of the mortgage borrowers to Experian, but this is not necessarily the property address, particularly for mortgages on non-owner occupied properties. Additional address information maintained within Experian's databases may prove useful in supplementing the repository addresses, as might historical information on borrower location. Nevertheless it is expected that such merges will be less accurate than those employing DII because the later are less reliant on address.

of the loan, culled from multiple sources, as well as vectors of dynamic data, such as the monthly performance of the loan from origination to termination, changes to its interest rate in each month (if a variable rate loan), and the associated loan balances. It should be noted that information from external databases is only used to supplement information about sample loans, not to add new loans to the sample. The NMDB sample frame will continue to be that established in the Experian data files. All information on mortgage performance will likewise come from Experian.

7.0   Evaluating the NMDB Sample Frame

A complete evaluation of the NMDB sampling frame may not be possible until the database is fully developed. However, at this stage of development the NMDB can be compared with HMDA data as alternative estimates of the U.S. mortgage origination market. Table 3 compares estimates of national quarterly origination totals from HMDA data and the NMDB. Loans are divided into two groups, based on whether they exceeded $75,000 in real 2013 dollars. Smaller loans are separated out because HMDA did not differentiate between first and second liens (which are generally small) prior to 2004.

The two databases track each other remarkably well, with HMDA totals slightly below those of the NMDB. This may stem from known gaps in HMDA's coverage. Loan originators that are very small or that operate exclusively in rural areas are exempt from HMDA's reporting, so their lending activity is not included in the HMDA data. Additionally, HMDA excludes commercial loans and (non-purchase) loans backed by properties that were previously mortgage-free. Many of these loans, however, may not be reported to the credit repositories either. For example, loans to corporations, loans made as part of a seller-financed property sale, and loans made by non-traditional lenders are unlikely to be in either database. Moreover, some types of loans may be missed by the NMDB though they are captured in HMDA. Lenders that retain all of their loans in portfolio, particularly credit unions, are known not to report their loans to the credit repositories, but are nevertheless still subject to HMDA reporting.

Table 3

Quarterly Loan Originations (1,000s of loans)

Quarter	$75,000 and Under NMDB[*]	HMDA	Over $75,000 NMDB[*]	HMDA	NMDB/HMDA (Percent Ratio) ≤ $75,000	> $75,000
1998-1	616	696	1,955	1,950	88.6 %	100.3 %
1998-2	755	893	2,278	2,247	84.6	101.4
1998-3	710	835	2,229	2,179	85.0	102.3
1998-4	705	816	2,808	2,676	86.3	105.0
1999-1	601	736	2,263	2,122	81.7	106.6
1999-2	661	862	2,143	2,046	76.7	104.7
1999-3	622	776	1,750	1,668	80.1	104.9
1999-4	544	669	1,418	1,349	81.3	105.2
2000-1	490	619	1,197	1,162	79.2	103.0
2000-2	571	779	1,470	1,466	73.3	100.3
2000-3	525	698	1,470	1,427	75.2	103.0
2000-4	473	608	1,451	1,384	77.7	104.8
2001-1	463	607	1,995	1,922	76.3	103.8
2001-2	610	872	2,814	2,789	69.9	100.9
2001-3	557	769	2,633	2,542	72.5	103.6
2001-4	574	757	3,558	3,434	75.8	103.6
2002-1	507	665	2,909	2,791	76.2	104.2
2002-2	504	702	2,541	2,521	71.7	100.8
2002-3	507	699	3,573	3,484	72.5	102.5
2002-4	540	717	4,778	4,608	75.2	103.7
2003-1	514	678	4,443	4,326	75.8	102.7
2003-2	630	889	5,579	5,443	70.9	102.5
2003-3	650	911	5,624	5,577	71.4	100.8
2003-4	430	634	2,966	2,989	67.8	99.2
2004-1	375	314	2,873	2,812	119.5	102.2
2004-2	460	392	3,518	3,538	117.4	99.5
2004-3	399	318	2,747	2,734	125.3	100.5
2004-4	343	299	2,816	2,798	114.7	100.6
2005-1	317	266	2,611	2,537	119.1	102.9
2005-2	389	310	3,137	3,057	125.6	102.6
2005-3	394	317	3,443	3,346	124.2	102.9
2005-4	321	273	2,832	2,760	117.4	102.6

Quarterly Loan Originations (1,000s of loans)

Quarter	$75,000 and Under		Over $75,000		NMDB/HMDA (Percent Ratio)	
	NMDB[*]	HMDA	NMDB[*]	HMDA	≤ $75,000	> $75,000
2006-1	287	245	2,370	2,284	116.8 %	103.8 %
2006-2	373	289	2,729	2,601	129.2	104.9
2006-3	360	273	2,582	2,450	131.9	105.4
2006-4	282	232	2,539	2,397	121.6	105.9
2007-1	251	210	2,240	1,997	119.6	112.2
2007-2	322	245	2,402	2,221	131.6	108.1
2007-3	291	222	1,929	1,848	131.0	104.4
2007-4	231	198	1,692	1,642	116.7	103.0
2008-1	215	195	1,874	1,759	110.0	106.6
2008-2	251	220	1,762	1,695	114.3	104.0
2008-3	217	191	1,283	1,254	113.8	102.3
2008-4	153	142	1,152	1,098	107.7	104.9
2009-1	161	156	2,126	2,017	103.1	105.4
2009-2	217	197	2,542	2,444	110.4	104.0
2009-3	200	183	1,786	1,750	109.0	102.0
2009-4	173	167	1,761	1,695	103.8	103.9
2010-1	137	135	1,315	1,291	101.2	101.9
2010-2	189	176	1,551	1,524	107.7	101.8
2010-3	179	176	1,921	1,871	101.8	102.7
2010-4	195	195	2,235	2,192	100.2	101.9
2011-1	158	166	1,387	1,376	95.3	100.8
2011-2	190	186	1,253	1,249	102.0	100.3
2011-3	201	203	1,544	1,533	98.8	100.7
2011-4	205	217	1,901	1,885	94.7	100.8
2012-1	199	207	1,846	1,828	96.4	101.0
2012-2	231	235	2,031	2,032	98.1	99.9
2012-3	241	249	2,310	2,279	96.8	101.4
2012-4	244	258	2,431	2,389	94.5	101.8
2013-1	238	251	2,189	2,173	94.8	100.7
2013-2	263	272	2,271	2,264	96.8	100.3
2013-3	242	247	1,753	1,748	98.0	100.3
2013-4	175	185	1,201	1,216	94.5	98.7

*National estimate based on 1-in-20 sample.

Appendix

The primary sources explored were the Home Mortgage Disclosure Act (HMDA) data, the Federal Reserve Bank of New York's Equifax Consumer Credit Panel, the CoreLogic property database, the servicing databases owned by CoreLogic and Black Knight Financial Services, and data available from the three national credit repositories—Experian, Equifax, and TransUnion. Public survey databases, particularly the American Housing Survey (AHS), were also considered. All of these sources share several desirable features such as: (1) the databases are de-identified containing no direct-identifying information such as borrower name, address, or Social Security number; (2) they are collected for other purposes, thus their use entails no new data collection from lenders, servicers or borrowers; and (3) all of them have been collected for a period of time and are expected to continue into the future.

However, each was also found to be deficient in significant ways.

The HMDA data include loan applications and underwriting outcomes for most mortgages with selected information about the loan, property, and borrower. The data are arguably the most representative publicly available existing data source about the mortgage market. However, the HMDA data contain no information on loan performance, little information on borrower credit-worthiness, and have up to a 21-month delay in release. The CoreLogic property database suffers from similar deficiencies. Although it has widespread coverage, the database contains very limited information on mortgage characteristics or performance and nothing on the borrower.

The Federal Reserve Bank of New York's Equifax Consumer Credit Panel provides a nationally representative 1-in-20 sample of individuals with credit records, observed quarterly from 1999 onward. However, mortgage loans are often represented by duplicate trade lines and important information is missing, such as loan purpose, owner-occupancy, pricing, loan-to-value ratio, income, and borrower demographics. Finally, these data are accessible at present only to the Federal Reserve System.

CoreLogic and Black Knight Financial Services produce loan-level databases with performance information collected from mortgage servicers. The servicing fields available from CoreLogic and Black Knight are relatively comprehensive in both variables and coverage: the CoreLogic database claims about 32 million active mortgage loans, while the Black Knight database claims about 31 million active mortgage loans. However, these data offer no assurance of being representative, as data are only collected from about 25 servicers each. Moreover, mortgages cannot be tracked if servicing is transferred. Other drawbacks include minimal borrower demographics and no information on other borrower's obligations.

The semi-annual AHS contains comprehensive information on a nationally representative 1-in-2,000 sample of mortgages of owner-occupied properties with very good information about the property and borrower demographic. However, the AHS has only limited information about the mortgage itself. As with the other nationally representative consumer survey data sources, AHS contains no information on mortgage performance, provides only a small number of observations, and is released with a significant lag.

14

The credit repository data from Equifax, Experian, and TransUnion are rich in credit information. By construction they incorporate data on credit card debt, installment loans, credit inquiries, and public records for the consumers they have in their respective databases. Their data can be linked to marketing datasets that provide borrower characteristics including age, gender, and marital status which, if validated, could be of potential use in a dataset. The credit repositories also maintain data on borrowers' changes of address and broader geographic classifications, such as the census tract. However, there are important areas that are not covered. They lack some information on borrowers (*e.g.*, race/ethnicity and income), mortgages (*e.g.*, loan product and contract rate), and the underlying property (*e.g.*, location and value).

Given the foregoing, FHFA and CFPB, along with other organizations most notably HUD, the Federal Reserve Board and Freddie Mac, decided that a modified derivative of the credit repository data offered the best source from which to construct a nationally representative comprehensive mortgage database. The three credit repositories all actively pursue loan servicers as data providers. As a result, they obtain information on almost the entire population of non-private mortgage loans made in the United States. Furthermore, they archive their data, making it possible to "jump start" the data collection process by going back in time, collecting data in almost the same fashion as if it had taken place in real time.

As part of the exploratory process, using a competitive procurement process, Experian was engaged by Freddie Mac to construct a prototype to confirm the appropriateness of using credit repository data for the database. This effort confirmed the concept but suggested that a number of steps needed to be taken in order to meet the design objectives.

First, it was recommended that the database should be a sample rather than a universal registry of loans. Second, while these data contain detailed information on loan performance and other borrower credit obligations, they are missing critical data items needed for the database such as the location and features of the property, demographics, and loans characteristics such as whether the loan had an adjustable- or fixed-rate mortgage and whether the loan was a refinance or for a home purchase. Thus, it would be necessary to access other data sources and merge information gleaned from them with the repository data in order to make the database comprehensive. Pilot testing also confirmed that the best method of merging data would rely on third-party blind matching conducted behind a firewall at the credit repositories.