NMDB®
National Mortgage Database

TECHNICAL REPORT 1

National Mortgage Database Technical Documentation

2021
January 21

This document was prepared by Robert B. Avery, Tim Critchfield, Craig Davis, Ian H. Keith, Ismail E. Mohamed, Saty Patrabansh, Jay D. Schultz, and Rebecca Sullivan. It is updated from the version published on March 10, 2020.

Table of Contents

1. Introduction

The National Mortgage Database (NMDB®) program is jointly funded and managed by the Federal Housing Finance Agency (FHFA) and the Consumer Financial Protection Bureau (CFPB). The program 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 three primary components:

(1) the National Mortgage Database (NMDB);
(2) the National Survey of Mortgage Originations (NSMO); and
(3) the American Survey of Mortgage Borrowers (ASMB).

The NMDB program enables 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 (HERA).[1] Specifically, FHFA must, through a monthly 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.[2]

For CFPB, the NMDB program supports policymaking and research efforts and helps identify and understand emerging mortgage and housing market trends. CFPB uses NMDB, among other purposes, in support of the market monitoring called for by the Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank Act), including understanding how mortgage debt affects consumers and for retrospective rule review required by this statute.[3]

In seeking to meet these objectives, FHFA and CFPB considered using existing databases but determined that none was sufficient, and that a new database, NMDB, had to be created.[4] NMDB is a de-identified loan-level database of closed-end first-lien residential mortgages. NMDB has the following features:

(1) it is representative of the market as a whole;
(2) it contains detailed, loan-level information on the terms and performance of mortgages, as well as characteristics of the associated borrowers and properties;
(3) it is continually updated;
(4) it has a historical component dating back before the financial crisis of 2008; and
(5) it provides a sampling frame for NSMO and ASMB.[5]

The core data in NMDB represent a statistically valid 1-in-20 random sample of all closed-end first-lien mortgages in the files of Experian, one of the three national credit bureaus.[6] When the

[1] Housing and Economic Recovery Act of 2008, Pub. L. 110–289, 122 Stat. 2654 (2008).
[2] FHFA interprets the NMDB program, including NSMO, as the "survey" required by the Safety and Soundness Act. The statutory requirement is for a monthly survey. Core inputs to NMDB, such as a regular refresh of credit-repository data, occur monthly, though NSMO is conducted quarterly.
[3] Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, 124 Stat. 1376 (2010).
[4] Please see the Appendix A for a discussion of sources available at the genesis of the program and their limitations.
[5] See NSMO Technical Documentation at http://www.fhfa.gov/nmdb.
[6] Experian was chosen through a competitive procurement process to assist in creating NMDB.

NMDB program began, an initial sample was drawn from all mortgage files outstanding at any point from January 1998 through June 2012. Since then the sample has been updated on a quarterly basis with mortgages newly reported to Experian. Mortgages (and their borrowers) are tracked in NMDB from at least one year prior to origination to one year after termination of the mortgage, whether that termination is through prepayment, adverse termination, or maturity.

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 the 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. The restriction of the NMDB frame to closed-end loans was made for two reasons. First, it mimics the reporting requirements of the Home Mortgage Disclosure Act (HMDA) and second, it reflects the practical fact that administrative data, which is a critical input for the NMDB data, is available for very few open-ended loans.

A random 1-in-20 sample of mortgages newly reported to Experian is added each quarter. 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 directly 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, detail how administrative data are merged into the NMDB, and provide the details of the current production version of the database (NMDB 12.0 as of this writing). The final section then evaluates the NMDB sample frame.

2. 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.[7] 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.[8] First, while Experian uses name, address and Social Security number for matching purposes only, this information is never 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.[9] Third, all access to the NMDB must be through a server at FHFA or CFPB and strictly controlled. Fourth, the NMDB – which is a sample and designed to describe the market as a whole – cannot be used for enforcement against any specific servicer or lender.

[7] A 10-year extension of this contract was signed in September 2017.
[8] 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.
[9] The Experian 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. At present, employees from Fannie Mae, Freddie Mac and nine Federal agencies have been granted access.

3. Selecting the Initial Sample

The credit repository core of the NMDB was 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 (as 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 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.

Since Experian retains archives of their data for 10 years or longer, data on mortgages that have been purged from Experian's current files due to FCRA 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 a sample from the December 2005 archive that captured loans that may have been purged by June 2012 and a sample from the July 2001 archive that captured loans that may have been purged by December 2005.

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 (currently the Metro 2® format, introduced in 1997 and made mandatory in 2018). 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.

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:

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

home-improvement loan); or

- trade lines reported by servicers with “Kind of Business Code” 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).

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 subject to any open-ended or revolving loans otherwise meeting one of the criteria being excluded. No other restrictions or criteria 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 “Enhanced Account Type Code” and “Kind of Business Code” criteria used for the June 2012 archive and 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 also 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 “Enhanced Account Type Code” and “Kind of Business Code” criteria used for the June 2012 archive and had “Account Balance Dates” of January 1998 or later and were opened in 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 all trade lines from the 2001 sample that should also be present in the 2005 archive.

4. Processing the Initial Sample

Once loans were selected for the initial NMDB sample, data from the July 2001, December 2005, and June 2012 archives was assembled. 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 servicer and loan identifiers as well as a de-identified borrower identification number called a PIN.[10] These could be used (imperfectly) to link TIP files to other account-level files both within an archive and over time.

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 overrepresented 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.

[10] The encrypted servicer identification and loan numbers are unique to the NMDB and are used 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 PINs are also unique to the NMDB and are randomized. Experian, however, maintains the mapping between the borrower (and servicer and loan) identification numbers used in their system and the PINs supplied to the NMDB team so that records in the NMDB associated with the same PIN will be associated with the same borrower ID in the Experian records.

Once the initial samples were de-duped, it was necessary to link archival records over time to create a composite picture of each sample loan (this is particularly important for loan performance as described in Section 7 and Appendix B). 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 was that PINs for a given borrower can change over time. There are times when a loan is first reported to the credit repositories that it 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 "PIN-merge" occurs. However, historical archives are stored with the PINs at the time of the archive. Thus, to properly connect borrowers (and mortgages) over time, it was necessary for Experian to provide a PIN-merge transformation table to map historical to current PINs.

As shown in Table 1, the de-duping process substantially reduced the size of the original NMDB sample. About 16% of the mortgage trade lines originally sampled from the June 2012 archive, 30% of the selections from the 2005 archive, and almost three-quarters of the selections from the 2001 archive were dropped.[11] 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 FCRA purge rules. However, many of these loans subsequently became current and could be found in later archives.

Table 1. Percent of Tradelines Dropped in Initial Sample by Archive Date

Archive Date	Sample Tradelines	Final Loans	Final Borrowers	Percent of Tradelines Dropped
Jul-2001	302,398	79,595	123,052	73.7
Dec-2005	2,955,675	2,057,651	3,362,299	30.4
Jun-2012	9,225,304	7,693,022	12,010,606	16.6

[11] A small percentage of the trade lines (1-2%) were dropped for other reasons including (1) loans with no balance or terms information; (2) those in American territories other than Puerto Rico, Guam or the Virgin Islands; (3) "frag files" missing information on all other consumer obligations; and (4) those with no information about the borrowers.

5. Updating the Sample

Under the NMDB sample design, credit records for borrowers associated with sampled mortgages are to be collected quarterly until one year after the mortgage is reported as closed. 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 from June 2012 onward.

The initial update of the NMDB from the June 2013 archive covered a full year of newly-reported mortgages since June 2012. From that date until June 2020, updates took place quarterly drawing from the last archive (either Wednesday or Saturday whichever is later) of the quarter (March, June, September and December). In July 2020, the archive date was shifted about two weeks into the next quarter to ensure that all (or almost all) loans reported within the quarter are captured.[12] Each quarterly update follows the same pattern. A 1-in-20 random sample of new 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, then only the line with the latest "Account Balance Date" is kept. 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 new loan is dropped.

Existing sample loans are also updated each quarter. Prior to the update, the PIN-merge transformation table is updated to account for "newly merged" PINs. To ensure that lagged information for all PINs newly added to the dataset is collected, the year-old archive is drawn each quarter for all active PINs for which this archive had not previously been collected. At present, upwards of 100,000 new loans are added to the NMDB each quarter (see Table 2). The number of mortgages ultimately added to the database is typically between 40,000 and 60,000 less than the number of raw trade lines originally selected for the update sample.

Over time, several additional "sample-cleaning" metrics have been added. Each quarter, a small number of old but newly reported mortgages are included in the Experian update sample. This can occur when a previously missing servicer decides to report to Experian or when servicing for a seasoned loan is transferred. These loans are disproportionately delinquent so it may also occur when a servicer only reports negative information. Believing these to be non-representative loans, it was decided to drop all loans more than ten years old when first reported. In addition, a small number of new loans are identified as: (1) closed in the same month as they are opened; (2) having steadily increasing balances (probably reverse annuity mortgages); (3) associated with persons for whom there are no other trade lines (frags); (4) associated with persons with no demographic or address data; or (5) added to Experian but removed in the next archive (probably an error in reporting). In each of these situations the loans are only

[12] Since the beginning of 2020 a partial update is done monthly for active sample mortgages. This allows the database to provide high-frequency information on mortgage performance. The monthly archive is drawn on the earliest Wednesday or Saturday between the 12th and 15th of the month.

provisionally added to the NMDB and are removed if the problem is not corrected within a year.

Table 2. Percent of Tradelines Dropped in Quarterly Updates by Archive Date				
Archive Date	Sample Tradelines	Final Loans	Final Borrowers	Percent of Tradelines Dropped
Jun-2013	648,224	493,211	766,557	23.9
Sep-2013	240,001	130,385	198,804	45.7
Dec-2013	174,404	101,401	151,842	41.9
Mar-2014	111,928	53,222	79,009	52.4
Jun-2014	146,406	76,943	113,819	47.4
Sep-2014	124,389	74,869	111,252	39.8
Dec-2014	124,323	75,586	111,920	39.2
Mar-2015	104,613	70,481	104,771	32.6
Jun-2015	129,737	90,732	135,661	30.1
Sep-2015	150,399	97,711	145,466	35.0
Dec-2015	124,413	87,383	129,524	29.8
Mar-2016	123,438	74,054	109,154	40.0
Jun-2016	111,797	82,449	121,914	26.3
Sep-2016	135,699	103,013	151,793	24.1
Dec-2016	177,386	107,620	159,614	39.3
Mar-2017	127,917	94,319	139,714	26.3
Jun-2017	129,953	80,382	117,742	38.1
Sep-2017	125,278	90,728	133,517	27.6
Dec-2017	154,468	90,764	133,148	41.2
Mar-2018	109,340	78,843	114,910	27.9
Jun-2018	117,964	79,660	115,981	32.5
Sep-2018	126,255	84,975	124,364	32.7
Dec-2018	134,045	76,775	111,886	42.7
Mar-2019	147,057	64,354	92,987	56.2
Jun-2019	189,499	78,661	114,530	58.5
Sep-2019	170,330	104,384	153,298	38.7
Dec-2019	168,428	123,795	182,332	26.5
Mar-2020	169,300	109,394	159,671	35.4
Jul-2020	221,233	181,873	269,431	17.8
Oct-2020	208,174	176,838	262,429	15.1

6. Merging with other Data Sources

Although Experian's archive is extensive, it does 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. 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 obtains much of the missing information from matches to administrative file records. The core administrative files come from Fannie Mae and Freddie Mac (the Enterprises), the Federal Housing Administration (FHA), the U.S. Department of Veterans Affairs (VA), and the Rural Housing Service (RHS). 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 is to use information about the borrowers, such as their names, Social Security numbers, addresses, and dates of birth. Using personally identifying information (PII), however, heightens 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 uses 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, FHA, VA, RHS, nor the Enterprises can receive PII from Experian. Second, Experian cannot access FHA, VA, RHS or Enterprise administrative data and borrower PII in the same place. Third, FHFA must not be able to match loans in the NMDB records to the specific administrative records from FHA, VA, RHS, or the Enterprises.

In December 2014, a process was initiated to supplement the NMDB data with administrative data from Fannie Mae and Freddie Mac. Subsequently, the same process was repeated with FHA, VA, and RHS. The process for matching the data follows seven steps:

(1) The data partner (*e.g.*, Fannie Mae) creates a unique anonymized identifier (AID) for each loan. This identifier, along with the borrower-level PII associated with each loan (including name, address, Social Security number, and date of birth), is transmitted directly to Experian using a secure portal. FHFA does not receive this information. Other administrative data on these loans is not sent to Experian.

(2) The data partner simultaneously sends 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 PII is included in the information sent to the FHFA data processing unit. The FHFA data processing unit transmits the administrative data along

with the associated AID to another unit within Experian which is separate from the unit that receives the PII in step (1).

(3) Behind a secure firewall to protect FCRA-regulated data, Experian matches the PII it receives from the data partner to the PII maintained in its own files on the borrowers in the NMDB to determine potential matches. When a potential match is identified, Experian compiles the PIN for each matched borrower.

(4) For all potential matches, Experian transfers the partner-supplied AID and the matched NMDB borrower PINs to a separate unit within Experian that has no access to the credit repository data or any PII. This is the same unit that received the administrative data from FHFA in step (2).

(5) The second Experian unit matches the AIDs received from the first Experian unit in step (4) with the AIDs sent by FHFA in step (2). For all matches, the second Experian unit forwards the administrative data they received from the data processing unit at FHFA, plus the matched borrower PIN that they received from the first Experian unit, to the NMDB development team at FHFA. The information sent to the NMDB development team includes neither the Enterprise-created AID nor any PII.

(6) The NMDB team compares the characteristics of the loans associated with the PINs received from the second Experian unit to the administrative information on the loans. If the information from both sources is consistent, the match is confirmed. The confirmation process also involves resolving issues when an NMDB loan is matched to multiple administrative loans or an administrative loan is matched to multiple NMDB loans. In these instances, the match with the closer fit is generally selected. The only instance where multiple matches are allowed are loans matched to either Freddie Mac or Fannie Mae and one of the government programs. These matches are allowed only if the codes for Freddie Mac or Fannie Mae indicate that their loan is government insured.

(7) A list of confirmed matches is sent to Experian. Upon confirmation, Experian stores the property address supplied as part of the PII file from the Enterprises but otherwise permanently destroys all PII used in the match.

The file matching process has been completed for all historical loans and is repeated quarterly for loans which have been newly added to the NMDB.[13] Generally, the matching process operates with a lag, such that it takes about 10 weeks after the end of a quarter for loans reported to Experian during that quarter to be flagged as confirmed matches. As is noted below, the matching process has proved to be very accurate. Match rates for all five administrative partners are in the high 90 percent levels (see Section 8). The only potential significant shortfall is for

[13] The administrative file matching process is done with all mortgages tracked in the NMDB. Because the database contains the complete credit bureau records of all sample loan borrowers perforce it will contain records on all mortgage loans associated with sample loan borrowers as long as they were closed within eleven years of the archive the sample loan was first reported or opened within a year of the sample loan being closed. This includes HELOCs and second liens as well as other closed-end first liens. Information on these loans is subject to the same NMDB cleaning and de-duping process used for sample loans. At present there are 55 million total mortgages tracked in the database of which only 13.1 million are sample loans.

loans on non-owner-occupied properties, whose match rates are closer to 80 percent. One explanation is that borrowers on some of these loans are partnerships or LLCs and thus not in the Experian files.

Efforts are currently underway to merge property record information into the NMDB, using similar third-party blind matching techniques. At present, all loans in the NMDB database added through mid-2020 have been matched to a property vendor's property records, including both transaction records and assessment files, with plans to match the remaining ones shortly.[14] For those loans where property matching is complete, data from the property vendor's servicing and private-label mortgage-backed securities (PLMBS) databases have also been matched which provides missing data elements for many of the non-government-affiliated loans in the NMDB. These match files are updated quarterly.

In addition, loans in the NMDB have been matched to loan-level information on loans reported under the Home Mortgage Disclosure Act (HMDA) from 1990 to 2019, loans included in the McDash servicing database, and to data files on the 2.1 million loans purchased by the Federal Home Loan Banks (FHLB). These matches do not involve PII and thus rely on less accurate matching techniques.[15] The HMDA matching is conducted once a year using a proprietary regulatory file which contains more granular loan information than the publicly released file, particularly with respect to dates and the original loan amount. The McDash and FHLB servicing files are updated and matched to the NMDB once a quarter.

Now in full production, the NMDB combines data from all these sources into a common file with one record per sampled loan. The record contains variables reflecting all the static characteristics of the loan collected from multiple sources, as well as vectors of dynamic data such as the monthly performance of the loan from origination to termination, the associated loan balances, and borrower credit scores and performance on other credit obligations. 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 continues to be that established in the Experian data files. Further, almost all dynamic information, such as loan performance, payments, loan balances, credit scores, and performance on other credit obligations is derived entirely from the Experian data files.

[14] 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 PII used in the match is discarded once the matching process is completed.

[15] Such merges use information other than PII common to the NMDB and the external dataset to perform a match. Primarily the matches 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 sometimes proves useful in supplementing the repository addresses, as does historical information on borrower location. Nevertheless, such merges are less accurate than those employing PII because the latter are less reliant on address.

7. Production

On October 30, 2017, NMDB 1.0 was certified and released for internal production use at FHFA and CFPB. The NMDB 1.0 dataset featured payment performance data for the 11.9 million NMDB sample mortgage loans and 18.6 million borrowers associated with those mortgages as of June 2017.[16] The initial release was used by the CFPB to support a public release of monthly county-level mortgage performance statistics. More detailed background information is provided on this component of NMDB 1.0 in Appendix B.

An updated production database, NMDB 2.0, was certified and released for internal production use at FHFA and CFPB on May 31, 2018. This process has been repeated quarterly through the present (NMDB 12.0 released on December 2, 2020).

NMDB 12.0 is a fully developed database consisting of information on 13.1 million sample mortgage loans and 20.3 million associated borrowers. The database is cumulative, in that updated information is potentially added on all sample loans for each quarterly update. Only about 20 percent of the database sample loans are currently active, but information on closed loans can still be updated due to new administrative file matching.

The database fully reflects the results of the NMDB matching process. The database is "complete" with no missing variable values for any sample loans. Most values reflect "hard" information from administrative or Experian files. However, no administrative matches could be made in some instances, particularly with older loans or loans not associated with one of the NMDB data partners. In these instances, information that would have been available from an administrative match, such as loan-to-value (LTV) ratio or borrower income, is imputed using statistical techniques. These techniques rely heavily on census-tract-level variables created from HMDA data or other administrative sources that reflect summary statistics of all mortgage loans made in the borrower's neighborhood during the same year that the loan in question was originated.

The dataset includes the state, county and census tract that the property associated with the loan is estimated to be in 2019 and in the year the loan was originated. Accompanying this is a set of tract-level information, which allows, for example, a user to determine if the loan would have qualified for Community Reinvestment Act (CRA) or Enterprise-goals preference or would have been conformable.

The dataset also includes "static" loan variables based on origination values contained in the Experian files including original term (months), loan amount, balloon status, origination date, and number of borrowers. Administrative matches are used to populate a number of other origination variables including property (collateral) value (used to compute LTV), income relied

[16] The production database excludes a small number of loan/borrower records because the borrower was not originally associated with the loan when it was sampled but added by Experian (or the servicer) later. This can only be picked up if the borrower happened to be in the NMDB because of another loan and thus will not be representative of all new borrowers added to loans. Consequently, they are excluded, and the production database should be considered a dataset of all loan/borrower combinations where the borrower was associated with the loan from the beginning.

upon in underwriting, back-end debt-to-income (DTI) ratio, occupancy status, ARM status, loan type (FHA, VA, RHS), ever purchased by a GSE (Fannie Mae, Freddie Mac, FHLB), in PLMBS, number of units in the property, manufactured housing and HARP loan flags, and initial contract interest rate. The previous loan payoff balance (refinance only) and simultaneous new second lien (piggyback or HELOC) data are used to derive combined-loan-to-value (CLTV) ratio and "cash-out" status.

Demographic data on loan borrowers derived from variety of sources are used to construct variables for birth month and year (used to calculate age), gender, race, ethnicity, and VantageScore 3.0 at origination. Information on the date of the borrower's first mortgage in the credit bureau and whether they have ever had a VA loan are used to infer veterans and first-time homebuyer status.

The dataset also contains several "dynamic" variables which vary by time. These include the monthly and quarterly unpaid loan balance and loan performance metric (*e.g.*, 30 days past due), quarterly credit scores for each borrower, the quarterly FHFA house price index applicable to the each loan's county (can be used to estimate the loans current loan-to-value ratio), quarterly loan inquiry counts for borrowers associated with the loan (can be used to forecast prepayment), variables reflecting loan modifications, and the closing date (if applicable). Beginning in the 4th quarter of 2019, the dataset has included quarterly summary information on balances and performance of five different types of non-mortgage credit associated with the borrowers for each active sample loan (auto loans, credit cards, student loans, HELOCs and 2nd liens, and all other loans).

The exact variables included in each new production release are described in the NMDB codebook which is updated quarterly. Details on the data editing, cleaning and imputation process which is employed in each quarterly data update are given in Appendices C and D.

8. Evaluating the NMDB Sample Frame

The NMDB is intended to be representative of the mortgage market as a whole. One way of testing whether this goal was achieved is to compare loan totals implied by the NMDB with control totals obtained elsewhere. Comparisons were made of implied national quarterly mortgage originations derived from the NMDB with control totals obtained from the five NMDB partner agencies (FHA, VA, RHS, Fannie Mae and Freddie Mac) for the period 1998 to 2019. The comparisons for Fannie Mae and Freddie Mac were done separately for loans backed by owner-occupied properties and those that were not. Comparisons were also made with administrative loan control totals for loans acquired by the FHLBs and loans in PLMBS in the CoreLogic LP database (collected from PLMBS trustee filings and estimated to be over 90% of the PLMBS market). Finally, implied NMDB total market estimates for loans backed by owner-occupied properties and those that were not were compared with those derived from HMDA data.

The NMDB-derived estimates of loan originations for the five data partners tracks control totals remarkably well (table 3). Almost all the quarterly ratios of NMDB estimates to control totals are over 90% with the vast majority over 95%. The average quarterly ratios are 96% and 94% for Fannie Mae and Freddie Mac loans backed by owner-occupied properties, 100% for FHA and VA, and 98% for RHS. RHS match ratios are shown for loans in the guaranteed program only. Much of the lending in the early years of the RHS program was direct, not guaranteed, and these loans are likely under-represented in the NMDB frame. In the control totals 17% of the RHS loans are direct, yet in the NMDB only 6% are direct. The somewhat lower ratios for Fannie Mae and Freddie Mac reflect lower match rates for the early years when Social Security numbers were not used the in the matching process. The average quarterly ratios for the two GSEs are 91% and 90% respectively for non-owner-occupied associated loans and 99% and 98% respectively for loans on owner-occupied properties when the analysis is restricted to loans originated since 2008.

The NMDB also compares well with control totals for the FHLBs and PLMBS. Average quarterly ratios are 101% for the FHLB and 114% for PMBS. The figures are somewhat misleading as both programs have diminished volumes in recent years contributing to small cell variance in the ratios. Over 70% of the FHLB loans and 98% of the PLMBS loans were originated in the ten-year period 1998-2007 preceding the great recession versus the 12-year period following it. For example, average PLMBS quarterly ratio for the earlier period was only 95%. Given that the Corelogic LP database likely missed some PLMBS loans, the NMDB database also likely does as well.

NMDB also compares well with HMDA data. On average, quarterly NMDB estimated totals for loans backed by owner-occupied properties are 104% of those reported in HMDA data. This may stem from known gaps in HMDA data's coverage. Loan originators that are very small or that operate exclusively in rural areas are exempt from HMDA reporting requirements, so their lending activity is not included in the HMDA data. Additionally, HMDA data 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 the HMDA data. Some lenders that retain all 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 requirements.

NMDB loans backed by owner-occupied properties compare more favorably with HMDA data than those backed by investor and secondary homes (an average quarterly origination ratio of 81%). Some of this difference may be accounted for by the fact that some investor loans are done through partnerships and LLCs and thus may not end up being reported to the consumer credit bureaus which form the basis of the NMDB.

Care should be exercised in placing too much weight on the specific quarterly numbers in table 3. There are issues with timing where a shortfall in one quarter is often offset by overshooting in the next. Missing information on opening date is common, particularly in the control files which generally reflect the current servicer, not necessarily the originator (NMDB cleaning rules use the original opening date when servicing is transferred). When this happens, servicers often assume that the loan is opened two months prior to the date of the first payment. Reflecting this, when the Experian date differs from the servicing file date for matched loans (this happens about 20% of the time for Fannie Mae and Freddie Mac and 30% of the time for FHA, VA, and RHS), 47% of the servicing file loans have an opening date of the 17th of the month—an implausible result. Also, the reported opening date falls on a Saturday or Sunday for 12% of the matched partner servicing files versus only 4% for the comparable Experian loans (mortgages almost always close on a business day).

The success of the NMDB coverage reflects the quality of the matching process. For every product-type except RHS Direct, 98% or more of the matches have a dollar-perfect match between the "original loan amount" reported in Experian and the amount shown in administrative records. For Freddie Mac and Fannie Mae, the rate is above 99%. As discussed above, however, "date opened" matches at a much lower rates ranging from two-thirds to 80% (except for RHS Direct which is 97%). This may occur, in part, due to a relatively high rate of servicing transfer for loans in the five partner programs. Loans "held in portfolio" and not placed in PLMBS securities or sold to the FHLB banks rarely have servicing transfers.

Despite the fact the Experian maintains the billing address for the mortgage and the administrative records maintain the property address, over 90% of the matches found an exact address match for all but the loans backed by non-owner-occupied properties. For non-owner-occupied properties, address matches were only about 60%, potentially contributing to lower overall match rates for these types of loan, particularly in earlier years. Social Security number (SSN) matching is used for all data partners except RHS in the current matching process. For example, over 99% of Fannie Mae, Freddie Mac, FHA, and 97% of VA loans had an exact match with SSN in 2020. The use of SSN compensates for the lack of an address match. However, SSNs were not available for matching for any loans prior to 2000 and not available on a comprehensive basis until after the Great Recession.

Table 3. Percent Ratio of NMDB Estimates of Originations to Control Totals, by Quarter											
	Fannie Mae		Freddie Mac							HMDA	
	Owner-Occupied	Other	Owner-Occupied	Other	FHA	VA	RHS	FHLB	PLMBS	Owner-Occupied	Other
1998Q1	87.9	78.5	85.7	75.6	101.1	101.0	58.3	82.6	75.9	-	-
1998Q2	91.2	77.9	86.2	78.7	104.0	103.8	65.0	92.5	80.4	-	-
1998Q3	90.6	80.1	88.1	78.7	100.2	102.1	58.5	99.5	73.7	-	-
1998Q4	91.0	81.7	86.9	75.2	100.6	100.1	61.1	136.5	76.8	-	-
1999Q1	90.7	79.2	88.0	74.4	102.4	101.5	64.2	107.5	76.6	-	-
1999Q2	92.1	79.4	86.6	71.5	100.2	104.6	64.0	99.4	80.0	-	-
1999Q3	91.4	77.8	83.2	72.0	106.0	105.9	61.7	116.1	78.1	-	-
1999Q4	91.2	76.5	75.2	65.7	106.6	113.6	63.2	138.0	73.5	-	-
2000Q1	85.9	71.8	79.5	70.2	107.8	115.1	70.0	106.2	79.8	-	-
2000Q2	89.9	76.2	81.3	69.6	101.6	104.5	71.3	103.9	80.9	-	-
2000Q3	88.9	77.4	87.8	76.1	104.9	103.4	72.8	114.5	82.9	-	-
2000Q4	89.8	76.8	84.3	72.4	104.1	107.7	66.2	117.5	80.8	-	-
2001Q1	91.4	77.3	85.7	73.7	99.8	100.5	64.5	138.5	84.6	-	-
2001Q2	91.4	80.8	86.3	75.8	101.4	102.1	63.2	127.0	95.3	-	-
2001Q3	93.4	79.5	90.0	74.0	103.0	104.3	66.9	99.6	93.9	-	-
2001Q4	94.0	81.4	89.7	77.8	101.7	102.5	65.5	88.3	97.4	-	-
2002Q1	93.9	78.4	89.2	74.5	105.3	102.1	74.0	120.8	100.7	-	-
2002Q2	92.7	78.4	88.2	77.3	101.6	104.6	66.9	128.0	96.4	-	-
2002Q3	92.5	81.6	89.6	76.0	103.3	103.6	75.0	121.3	105.7	-	-
2002Q4	90.5	79.2	88.1	75.1	103.0	103.4	71.4	122.1	111.4	-	-
2003Q1	93.8	82.7	94.9	84.8	102.0	101.8	76.9	100.9	108.8	-	-
2003Q2	94.6	84.4	91.9	85.3	99.9	101.7	74.8	99.6	101.2	-	-
2003Q3	94.7	83.9	92.7	88.6	101.3	100.7	79.4	97.5	107.7	-	-
2003Q4	95.4	81.9	93.5	81.6	100.1	101.8	99.2	92.8	101.5	-	-
2004Q1	95.5	80.9	94.5	78.3	100.5	99.1	109.1	91.1	99.8	104.6	90.2
2004Q2	95.7	82.3	92.8	78.8	100.6	99.4	102.6	78.0	102.7	99.7	96.8
2004Q3	96.9	82.5	92.6	76.5	98.9	95.2	116.3	82.3	99.6	102.1	95.7
2004Q4	96.0	84.5	91.8	74.8	99.1	97.9	105.6	102.6	100.4	101.5	92.9
2005Q1	97.2	84.7	93.3	78.6	98.4	96.5	117.9	91.6	98.7	104.4	92.0
2005Q2	95.5	83.8	92.3	80.6	97.9	100.2	129.2	88.6	98.6	104.0	96.6
2005Q3	95.4	81.9	94.7	80.0	98.5	99.4	117.4	88.0	100.2	104.1	98.3
2005Q4	92.9	82.0	90.7	73.7	102.6	98.1	119.1	96.4	96.2	102.6	95.7
2006Q1	93.1	82.2	92.7	78.9	99.3	99.1	133.6	92.9	90.8	105.9	91.7
2006Q2	94.3	81.2	92.8	80.9	103.0	101.7	116.1	76.5	94.2	106.9	95.4
2006Q3	93.7	81.9	92.6	81.6	101.7	97.4	129.2	89.1	93.6	107.7	96.3
2006Q4	92.1	79.6	93.1	82.4	99.4	101.9	124.8	110.7	92.4	107.3	93.4
2007Q1	92.8	82.8	90.8	78.9	99.2	99.6	122.9	111.7	87.1	113.9	95.8
2007Q2	91.7	83.8	89.0	76.9	98.9	98.6	114.5	120.6	92.6	111.0	93.7
2007Q3	92.0	81.4	90.7	81.8	97.4	101.3	115.8	118.4	82.8	108.8	83.3
2007Q4	93.3	83.6	92.4	82.1	97.4	99.0	118.7	90.5	66.2	106.8	79.3

Table 3. Percent Ratio of NMDB Estimates of Originations to Control Totals, by Quarter (Cont.)											
	Fannie Mae		Freddie Mac							HMDA	
	Owner-Occupied	Other	Owner-Occupied	Other	FHA	VA	RHS	FHLB	PLMBS	Owner-Occupied	Other
2008Q1	93.6	84.2	95.5	83.6	98.9	102.7	114.5	94.3	59.0	110.5	76.8
2008Q2	92.8	82.5	95.6	80.5	99.5	98.9	112.6	90.2	59.2	108.8	67.7
2008Q3	92.3	87.4	94.3	81.3	98.7	100.9	105.2	89.4	59.4	108.1	65.2
2008Q4	91.0	87.6	101.0	84.1	98.5	100.4	99.1	85.1	71.9	109.7	66.3
2009Q1	95.8	85.9	98.7	88.5	100.9	99.2	112.9	94.0	84.1	108.5	69.8
2009Q2	96.8	91.5	97.6	92.6	100.7	99.3	99.2	89.6	64.7	106.2	74.9
2009Q3	97.6	84.7	97.4	90.0	98.9	95.7	106.1	100.0	84.2	104.3	70.4
2009Q4	99.7	87.0	100.7	87.7	99.9	98.3	103.8	106.7	92.2	106.2	73.8
2010Q1	99.9	91.0	100.2	89.1	99.0	99.9	103.3	82.4	107.8	104.2	73.1
2010Q2	100.7	95.3	98.2	95.5	99.9	99.7	123.1	80.7	105.8	104.8	75.6
2010Q3	99.6	94.0	98.5	91.1	99.3	98.0	102.1	82.9	129.2	105.7	79.0
2010Q4	100.2	90.6	97.9	89.5	99.2	99.4	109.6	98.9	206.7	103.3	79.5
2011Q1	100.2	92.4	98.7	88.4	100.9	100.3	100.1	90.6	252.7	102.0	77.9
2011Q2	99.9	91.5	98.4	90.4	101.1	99.7	103.4	83.6	136.4	103.6	77.6
2011Q3	98.9	91.1	99.9	93.0	100.1	100.7	103.8	80.2	155.9	104.2	79.0
2011Q4	99.5	90.9	95.9	90.0	99.5	100.2	105.4	91.9	208.1	102.0	79.3
2012Q1	99.0	96.7	95.7	93.2	100.3	99.6	98.7	91.7	188.4	103.1	88.7
2012Q2	99.5	96.1	97.1	95.3	101.9	100.5	103.1	87.3	146.7	101.3	92.5
2012Q3	100.2	99.6	99.0	96.7	100.2	99.9	105.5	84.8	149.6	102.4	95.5
2012Q4	100.7	93.5	98.8	91.6	99.4	101.2	101.6	101.2	167.4	102.9	87.0
2013Q1	100.8	93.5	96.9	92.1	101.6	100.4	97.8	97.2	146.9	102.3	87.3
2013Q2	100.2	95.3	96.3	90.5	100.3	101.3	99.1	91.3	150.2	101.1	92.6
2013Q3	100.6	98.0	94.3	87.1	100.3	99.3	106.1	96.4	203.0	100.0	91.0
2013Q4	100.2	93.7	94.7	88.2	96.5	95.9	98.5	83.6	257.0	100.2	83.6
2014Q1	96.3	92.6	95.6	87.9	99.7	97.8	101.6	86.7	232.1	97.9	78.3
2014Q2	102.9	96.6	99.8	92.4	100.1	101.1	106.3	92.1	185.5	102.2	85.2
2014Q3	101.6	93.6	95.3	91.9	98.3	97.1	104.3	83.6	181.0	99.7	84.7
2014Q4	100.7	93.6	94.0	84.7	98.6	96.6	103.4	97.0	116.2	99.6	81.5
2015Q1	99.6	90.6	93.2	87.1	100.3	98.0	106.3	110.1	105.5	101.9	82.8
2015Q2	101.4	93.7	96.2	87.1	100.1	96.9	110.8	92.0	117.2	100.9	85.5
2015Q3	101.8	92.1	97.7	85.1	97.9	97.6	102.4	95.5	75.9	100.1	83.4
2015Q4	100.1	90.2	96.2	85.7	96.7	95.9	106.0	92.7	241.3	100.3	78.5
2016Q1	97.7	87.8	97.7	86.3	98.9	100.3	110.6	98.2	250.9	100.8	78.9
2016Q2	98.1	90.7	96.3	87.9	98.4	97.8	105.8	101.6	228.3	99.6	88.2
2016Q3	97.6	87.9	98.0	89.4	97.8	98.2	106.5	102.8	198.2	100.3	86.1
2016Q4	97.1	90.6	97.3	89.8	98.5	97.1	103.3	106.0	171.3	100.0	82.9
2017Q1	98.1	88.4	97.7	82.9	97.7	96.9	106.5	104.9	133.3	100.2	74.9
2017Q2	98.0	87.0	98.2	92.1	97.9	96.8	101.9	113.8	126.1	100.5	79.4
2017Q3	97.3	86.8	97.8	91.1	98.4	96.0	102.7	105.7	126.1	99.9	78.5
2017Q4	98.1	84.7	99.1	83.2	99.6	97.9	101.4	102.8	101.4	101.0	72.7

Table 3. Percent Ratio of NMDB Estimates of Originations to Control Totals, by Quarter (Cont.)											
	Fannie Mae		Freddie Mac							HMDA	
	Owner-Occupied	Other	Owner-Occupied	Other	FHA	VA	RHS	FHLB	PLMBS	Owner-Occupied	Other
2018Q1	98.7	81.4	98.9	82.4	98.9	99.1	108.6	102.0	75.4	103.3	60.1
2018Q2	98.0	86.7	97.9	87.4	97.8	100.6	100.7	118.1	61.7	102.4	59.6
2018Q3	99.4	86.4	97.9	91.7	98.8	97.9	108.9	109.3	53.9	103.4	59.6
2018Q4	98.9	90.7	99.7	98.7	99.1	97.5	108.9	109.3	54.9	103.4	63.1
2019Q1	100.6	94.4	104.1	96.9	98.6	100.1	109.4	116.4	74.8	105.1	62.7
2019Q2	99.0	90.6	98.7	93.9	98.3	99.4	104.1	119.9	60.3	103.3	60.4
2019Q3	98.5	89.8	98.3	93.1	100.2	99.9	103.8	111.9	55.6	103.1	61.4
2019Q4	98.8	89.9	98.7	92.4	98.2	98.8	103.8	124.2	49.9	102.0	62.4

Another validation of the NMDB matching process is the high correlation between credit bureau codes indicating FHA, VA, RHS, Freddie Mac, and Fannie Mae loans and the match results (which do not depend upon these codes). Over 98% of FHA, 99% of VA and 82% of RHS loans matched to administrative files were correctly marked by bureau account type codes. Only two-thirds of the Freddie Mac and Fannie Mae loans were correctly identified despite rules that servicers are supposed to report these loans to the bureaus. Some of this may be due to timing, as the reporting rules were not put into effect until 2010. More than 80% of Freddie Mac and Fannie Mae loans originated since 2008, for example, are correctly flagged.

Overall, these results suggest that the NMDB should provide a very accurate representation of the mortgage market as a whole and for the market subprograms represented by FHA, VA, RHS, FHLB, and loans backed by owner-occupied properties sold to Fannie Mae and Freddie Mac, and in PLMBS. There is a somewhat lower level of accuracy in earlier years, but the difference is not substantial. Loans backed by investor and secondary properties do appear to have somewhat lower representation, but coverage levels still should be above 80% even in earlier years.

Appendix A. Origins of NMDB

Prior to deciding to develop the NMDB, FHFA and CFPB considered a number of alternatives to meet their data requirements. The primary alternatives explored were the Home Mortgage Disclosure Act (HMDA) data, the Federal Reserve Bank of New York's Equifax Consumer Credit Panel, the property and 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, and until 2018 little information on borrower creditworthiness, and have up to a 21-month delay in release. The CoreLogic and Black Knight property databases suffer from similar deficiencies. Although they have widespread coverage, the databases contain 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 (currently) from about 55 servicers each. Moreover, mortgages cannot be tracked if servicing is transferred. Other drawbacks include minimal borrower demographics and no information on the borrower's other obligations.

The biannual 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 demographics. 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.

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.*, 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 HUD, the Federal Reserve Board, Freddie Mac, and others, 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, that the database be restricted to closed-end first lien mortgages, mimicking the coverage of HMDA data and the availability of matching information. Third, 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.

Appendix B. Background on Mortgage Performance Reporting

Almost all closed-end first-lien mortgages, such as those in NMDB, have a payment-in-arrears structure. That is, the mortgage payment for a month (*e.g.*, January) is generally due on the first day of the next month (*e.g.*, February 1). Moreover, the first "ever" payment on a mortgage is generally due on the first day of the second full month after the mortgage closing date. For example, borrowers who close on their mortgage on January 15 will have their first payment due on March 1. These borrowers would have prepaid the interest for the period covering January 15 through January 31 at their mortgage closing. One component of each monthly mortgage payment is the interest of the previous month based on the balance at the beginning of the month. While the monthly mortgage payment is generally due on the first of the next month, most lenders allow a 15-day grace period for borrowers to pay. However, if the payment is not received by the 15th, the mortgage loan is considered past due 15 days. Thereafter, a loan not paid by the X-th date after the due date is considered X days past due.

Each month mortgage servicers report information to the credit bureaus for each mortgage loan they service as of a snapshot date (balance date). Generally, the bureaus will accept only one report per loan per month and will not accept a report unless it has performance information. Servicers report three measures of performance: (1) the account condition code which describes the condition of the mortgage, *e.g.*, whether it is open, paid in full, closed, transferred, or inactive; (2) a special comment code which provides special information on the mortgage such as a loan modification, location in a disaster area county, or dispute by the borrower; and (3) a loan status code which provides information on how many days "past due" a loan is as determined by the oldest non-paid payment (loan payments are generally applied against the oldest non-paid payment). Industry and Metro 2 credit bureau reporting guidelines (available since 1997) differentiate between loans that are current or past due 29 or fewer days; 30 to 59 days past due; 60 to 89 days past due; 90 to 119 days past due; 120 to 149 days past due; and 150 to 179 days past due; and 180 or more days past due. This is the classification used for most loans. If a loan becomes 90 days past due under many mortgage contracts the lender can declare the loan "in default." The borrower then typically has 90 days to become current. If not, the lender can file a foreclosure action, in which case the loan status is changed from "days past due" to some form of foreclosure or collection. If the borrower files for bankruptcy, the loan may be assigned a "bankruptcy" status even if the payments are current.

Under normal conditions the bureaus will not accept a report unless it contains information on the loan's status. However, a status code may be suppressed or not reported for some loans. This can occur for a variety of reasons—a borrower's payments may have been suspended because of a natural disaster; reports may not be supplied for the period between a loan's closing or the first due date; status is often not continually reported when loan servicing is sold from one lender to another when it takes time for the acquiring lender to set up reporting; status updates are often not reported for loans in foreclosure or other forms of serious delinquency (there are spikes in missing values for August 2015 and April 2016 because several larger servicers had problems with their servicing systems). The borrower may dispute a performance report, in which case it may be missing.

There are also loans which do not fit these circumstances—some loans do not have due dates of

the first; others have bi-weekly or quarterly payment requirements; some borrowers make partial payments (often to what is called a suspense account) which can leave them in a perpetually past-due status; others can make extra payments (curtailment) to reduce their loan balance more rapidly.

Servicers reporting to the credit bureaus using Metro 2 guidelines are supposed to follow the guidelines described above. Thus, a loan with a due date of January 1 will be considered 30 days past due on January 31 if the payment has not been received by that date (the "days past due" standard). In the past, however, many lenders used a "billing cycle month" standard. Under the "billing cycle month" standard, a loan was not considered "30 days" past due until the due date of the next month (*e.g.*, February 1 for a January 1 payment). The "billing cycle month" standard, associated with the older Metro 1 reporting format, was phased out over the 2000s for bureau reporting. However, this has not necessarily happened for other regulatory reporting. Mortgage delinquency metrics reported to the Federal Financial Institutions Examination Council (FFIEC) for banking institutions can be based on either "days past due" or "billing cycle month" standards at the reporter's discretion. Credit unions used the "billing cycle month" standard until 2013 when they were required to report using both methods. Freddie Mac and Fannie Mae report delinquency statistics for loans in their security pools using the "billing cycle month" standard.

Within the Metro 2 reporting guidelines and for other delinquency reporting there is also variation based on precisely when a loan's status is measured. Under the Mortgage Bankers Association's "MBA" method, a loan is considered past due X days if a payment is not received by close of business (COB) on the X-th day following its due date. That is, a loan with a due date of March 1 is considered 30 days past due at COB on March 31. Under the Office of Thrift Supervision's "OTS" method, a loan is considered past due X days if the payment has not been received by COB on the X+1-th day (*e.g.*, April 1 for a March 1 due date).

The credit bureaus allow loan servicers to choose whichever reporting day within the month that they wish to use, and either the MBA or OTS method. Currently, about 90% of reporters use the same day of the month every month and the same day for all of their loans. For NMDB loans active in 2014 and later, the modal report day (31%) was the last day of the month; 16% were on the 5th; 12% on the 7th, and 8% on the 21st. For "prime" first lien closed-end mortgages, which dominate the NMDB, lenders generally use the MBA method. Subprime servicers, however, who played a significant role in the 2003 to 2007 period, typically used the OTS method.

These differences in reporting day and method can lead to significant variation in the incidence of delinquency for loans with identical payment patterns when comparisons are made month-to-month or between lenders with different reporting patterns. This is shown in Table B-1. Servicers who report at the end of the month using the MBA method will maximize the 30-day delinquent count in the seven months with 31 days because the reporting day is the first day a loan can be 30 days delinquent. Lenders reporting in the latter half of the month—but not on the last day—will tend to systematically show lower delinquency rates.

These distinctions will matter when aggregated measures of delinquency are computed, particularly those for 30-days past due. For example, in 2016 lenders in the NMDB reporting on

the last day of the month showed an average 30-day delinquency rate 0.76 percentage points higher in the seven months with 31 days than they did for the five months with 30 days or less. For servicers reporting between the 16th and the second-to-last day of the month there is only a 0.03 percentage point difference.

These reporting differences can cause systematic differences across states as well. For example, 47% of the 2016 reporters in Mississippi were end-of-the-month reporters versus 17% in Alaska, almost surely influencing the number of 30-day delinquencies. Because reporters can only report loan status once a month and it is impossible to know when a loan payment was received for many loans, this bias is difficult to correct for. If no adjustment is made, users of a bureau-based delinquency series need to be cautious in making comparisons across geographic units. Also, as just noted, numbers for the NMDB are likely to show persistent monthly patterns if results are not seasonally adjusted.

The mixture of reporting patterns in the credit bureau data is likely to lead to systematic differences in aggregate delinquency metrics constructed from the NMDB data when compared to other delinquency measures. For example, The MBA National Delinquency Survey asks respondents to classify loans by their status at COB on the last day of the quarter using the MBA method although it appears that the lender can use either the "billing cycle month" or "days past due" standard. If the "days past due" standard is used it means that reports for March and December (which each have 31 days) will show persistently higher 30-day delinquency rates than those of June and September (which each have 30 days). Similarly, 90-day delinquency rates will be lower in the first quarter except for leap years. The degree of seasonality will depend on what percentage of the reporters use the "billing cycle month" versus "days past due" standard. FFIEC call report statistics, which are also reported COB on the last day of the quarter, will also exhibit seasonality depending on the mix of lenders using different methods. FFIEC statistics are further clouded by the fact that lenders can use either the MBA or OTS accounting method.

Delinquency statistics reported by Freddie Mac and Fannie Mae for loans in their security pools should show the least month-to-month distortions. Both companies use an end-of-month measure computed using the MBA and "billing cycle month" standard which should lead to stable monthly patterns. Given that the standard mortgage contract is based on a monthly payment standard there is a compelling argument that the "billing cycle month" standard for measuring delinquency is the most appropriate. Nevertheless, that is not the standard used by most credit bureau reporters and thus is not the standard reflected in the NMDB data.

The delinquency data in the NMDB are built from the lender reports supplied by Experian but with some additional processing. The performance information supplied by lenders for loan status, account condition, and special comments is static; that is, each month when the servicers update the performance data for a loan, the previous values for these variables are overwritten with new information. The values supplied in the previous months can only be recovered from archives. However, all the credit bureaus maintain an abbreviated record of historical performance, known as a payment grid, which is not overwritten, but can be (and is) updated. Under FCRA rules the payment grid can only go back 84 months. When an initial report is supplied for a month (say June 2016) the "June 2016" element of the payment grid is initially

populated. However, in subsequent filings the servicer can change the "June 2016" value. This can happen for a variety of reasons—the lender can catch an error, they may have inadvertently failed to report performance in the first filing, the consumer could dispute the report and get the record changed, or the report could subsequently be suppressed, for example, because the borrower was impacted by a natural disaster.

In general, the monthly performance measure in the NMDB is constructed from the payment grid, using the most recently reported information for a given month. Payment grids retrieved from archival data—which were collected quarterly from June 2012 on and semi-annually before that—were used to piece together a full measure of performance and to get around the 84-month limitation on current data. An additional problem is created when loans are transferred from one servicer to another. Here, payment grids need to be combined for two different reporters to create a continuous measure of performance. Often when this happens, the transferring servicer will initially report the loan as delinquent but then correct it when they receive the transfer notice. Transfers often create gaps in the payment grid when the new servicer is slow to report the loan. FCRA rules also place restrictions on how the new servicer reports performance under the assumption that some borrowers may have sent payments to the wrong place.

The effect of this process is that the initial performance report for a loan is often subsequently changed. On net, this tends to improve the overall measure of performance, but in recent years the change is small. For example, the initial NMDB report for June 2016 differed from "final" report in June 2017 for 1.3% of the cases. The majority of these were blanks in the initial report but there were some real changes. Changes went both ways—2.3% of the loans originally reported as 30-days past due were corrected to current. But an almost equivalent number were changed from current to delinquent.

On balance, the updating process reflected in the NMDB is likely to mean that in recent years delinquency measures in the NMDB will be slightly more positive than other indices, such as the Equifax index, which are compiled only from the initial report. However, during the mid-2000s when the private label subprime market was a significant part of the mortgage market, sale of servicing was more prevalent and more likely to have led to initially inaccurate delinquency reports. Here, the NMDB data show noticeable differences from indices based on initial reports.

Another difference arises when seriously delinquent loans are transferred within an organization (say from normal servicing to "work out" departments). It is not unusual for the loan to be reported as open and delinquent by both departments creating a double counting if not corrected. In constructing the NMDB these reports are combined but may not be in other indices which are based on open accounts with positive balances. Consequently, indices of serious delinquency constructed from the NMDB will likely be lower than those constructed from other sources.

There is some ambiguity as to how to define an open account. It is not unusual for lenders to initiate foreclosure actions on small mortgage loans but never complete the process, perhaps because they decide the property isn't worth acquiring. In other cases, state law allows lenders to maintain a claim on the borrower, termed a deficiency judgement, after a foreclosure. These loans can remain on the Experian files as open, with positive balances, for a long time until they are purged by FCRA rules. However, the borrower may well have lost title to the house or

moved out much earlier in the process. Currently, the NMDB team arbitrarily treat these accounts as closed in the NMDB after 24 months, but others may have different rules.

Finally, there are two recent developments which impact performance metrics. Traditionally, when a consumer entered the bankruptcy process all their loans were tagged with performance bankruptcy codes even in cases where the loan (often a mortgage) was not included in the bankruptcy and the consumer was continuing to pay on time. In 2019 the industry guidance was changed to recommend suppressing all credit reporting on these loans irrespective of whether the borrower was current.

Another development is the passage of the CARES Act in March 2020. The CARES Act provided the option for homeowners with federally backed or funded mortgages to request forbearance (a pause) of mortgage payments for up to 180 days. During this period, loan servicers could not take any action which had an adverse impact on the credit of a borrower. Effectively, this "froze" the status of the loan at its pre-forbearance level even in cases where the borrower was not making payments. Potentially, this could lead to artificially low levels of delinquency.

There are two monthly performance vectors contained in the NMDB. One is the "raw" data whereby months where a performance report for a loan was either suppressed or not given are tagged as "missing." The second metric, and the one published by the FHFA, invokes a "stale account rule." Here the most recent non-missing performance metric is substituted for all months in the raw data vector with missing data. The rule extends back three months for most loans but goes back up to two years for loans that are seriously past due (180 days or more or in bankruptcy, foreclosure or collection). This rule emulates how performance would be treated in constructing most credit scores.

Table B-1.

Impact of Reporting Cycle Standard and Reporting Method on Delinquency Measurement
Reporting of Days Past Due for a Mortgage where Payments were Stopped*

	MBA Method**							OTS Method**						
Month Payments Stopped	Current Month	Month Plus 1	Month Plus 2	Month Plus 3	Month Plus 4	Month Plus 5	Month Plus 6	Current Month	Month Plus 1	Month Plus 2	Month Plus 3	Month Plus 4	Month Plus 5	Month Plus 6
End of Month Reporters - Days Past Due Standard														
January	**D30**	**D30**	D60	D90	D150	D180	D180	C	D30	D60	D90	D120	*D150*	*D180*
February	C	D30	D60	D90	*D120*	*D180*	D180	C	D30	D60	D90	D120	*D150*	*D180*
March	D30	D60	D90	D120	D150	D180	D180	C	*D30*	*D90*	D120	D150	D180	D180
April	*C*	*D60*	D90	D120	D150	D180	D180	C	D30	*D60*	*D120*	D150	D180	D180
May	D30	D60	D90	D120	D150	D180	D180	C	*D30*	*D90*	D120	D150	D180	D180
June	*C*	*D60*	D90	D120	D150	D180	D180	C	*D30*	*D90*	D120	D150	D180	D180
July	D30	D60	D90	D120	D150	D180	D180	*C*	*D60*	D90	D120	D150	D180	D180
August	D30	D60	D90	D120	D150	D180	D180	C	*D30*	*D90*	D120	D150	D180	D180
September	*C*	*D60*	D90	D120	D150	D180	D180	C	D30	*D60*	*D120*	**D150**	***D150***	*D180*
October	D30	D60	D90	D120	D150	D180	D180	C	*D30*	*D90*	**D120**	**D120**	D180	D180
November	*C*	*D60*	**D90**	***D90***	*D150*	D180	D180	C	*D30*	***D90***	**D90**	D120	*D150*	*D180*
December	D30	**D60**	***D60***	*D120*	D150	D180	D180	*C*	***D60***	**D60**	D90	D120	D180	D180
Middle of Month Reporters - Days Past Due Standard														
All Months	C	D30	D60	D90	D120	D150	D180	C	D30	D60	D90	D120	D150	D180
Billing Cycle Month Standard Reporters**														
All Months	D30	D60	D90	D120	D150	D180	D180	C	D30	D60	D90	D120	D150	D180

Impact by Number of Days in the Month and Timing of Reporting Date
Reporting of Missed Payments and Subsequent Cure Using the Days Past Due Standard and MBA Method

	March (31 Day Month)							April (30 Day Month)					
	Performance Month							Performance Month					
Date Cured	March	April	May	June	July	August	Date Cured	April	May	June	July	August	Sept.
End of Month Reporters													
15-Apr	D30	C	C	C	C	C	15-May	C	C	C	C	C	C
15-May	D30	D60	C	C	C	C	15-Jun	C	D60	C	C	C	C
15-Jun	D30	D60	D90	C	C	C	15-Jul	C	D60	D90	C	C	C
15-Jul	D30	D60	D90	D120	C	C	15-Aug	C	D60	D90	D120	C	C
Report on 22nd of Month													
15-Apr	C	C	C	C	C	C	15-May	C	C	C	C	C	C
15-May	C	D30	C	C	C	C	15-Jun	C	D30	C	C	C	C
15-Jun	C	D30	D60	C	C	C	15-Jul	C	D30	D60	C	C	C
15-Jul	C	D30	D60	D90	C	C	15-Aug	C	D30	D60	D90	C	C
Report on 7th of Month													
15-Apr	C	D30	C	C	C	C	15-May	C	D30	C	C	C	C
15-May	C	D30	D60	C	C	C	15-Jun	C	D30	D60	C	C	C
15-Jun	C	D30	D60	D90	C	C	15-Jul	C	D30	D60	D90	C	C
15-Jul	C	D30	D60	D90	D120	C	15-Aug	C	D30	D60	D90	D120	C

Note: This illustration is for loans where the payment due date is on the first of the month.
C = Current, D30 = 30-59 days past due, D60 = 60-89 days past due, D90 = 90-119 days past due,
D120 = 120-149 days past due, D150 = 150-179 days past due, D180 = 180 or more days past due.

*__Bold__ indicates where performance reporting stays the same and *italics* indicates where reporting skips a reporting cycle.

**See text for explanation of "MBA" and "OTS" methods; and "Days Past Due" and "Billing Cycle Month" standards.

Appendix C. Cleaning and Editing Data in the NMDB

Constructing the NMDB is a process fraught with errors of omission and commission (*e.g.*, data entered incorrectly). It is built out of thousands of servicing databases maintained by individual servicers. Such databases are generally going to contain little inaccurate information but suffer from errors of omission whereby certain variables are simply not collected and others contain many missing values. The next step, when information from the 5,000 or so individual servicers is reported to and aggregated at Experian, creates an opportunity for more errors. Loans may be assigned to the wrong person. Servicers may apply Metro 2 reporting guidelines in an inconsistent manner. This may lead to loans which should be eligible for NMDB sampling being missed and to loans which should not have been eligible being included.

Further error is likely to be introduced when administrative loan information from the NMDB's five data partners is merged into the NMDB. The administrative units themselves may receive incomplete or erroneous data from servicers. Matching these loans to the NMDB may be imperfect, potentially leading to inconsistencies between the administrative-data-based values and credit bureau-based values for the same loan.

The NMDB team has no control over data cleaning decisions taken by Experian or the NMDB's data partners. However, many problems can emerge during the cleaning process at FHFA, which is in the NMDB team's control. The production team executes over 50 programs with over 100,000 lines of code to process a normal NMDB update cycle. A significant portion of this code is used to determine whether new loans should be added to the NMDB (as described in Section 5) or whether loans are appropriately matched to administrative data (as described in Section 6). Most of the remaining processing is devoted to determining the dataset values of key variables.

The final NMDB dataset consists primarily of data provided through the Experian credit files supplemented by administrative data for variables that are not available from Experian. When there is overlap between the two sources, generally the Experian value is used in the dataset. This is done under the belief that the credit bureau data are generally cleaner and more consistently collected than raw servicing data. With small exceptions described below, values of the key variables including original loan amount and term, date opened and closed, monthly loan performance, balance, and payments are all derived entirely from Experian.

Complete information on original loan amount and term and opening date are conditions for a loan being eligible for the NMDB. Nevertheless, in a few cases there is some ambiguity in these variables that needs to be addressed. This can happen when there is an inconsistency between Experian or administrative data or within the Experian data. Typically, these inconsistencies are resolved by choosing the oldest opening date or term or loan amount most aligned with balance information.

More serious is the problem of ambiguity on loan closing. About 3 percent of all NMDB loans end with an uncertain disposition—a servicing transfer, death of a borrower, foreclosure started but not finished, or just a stop in reporting. Several rules are employed to resolve these cases. If a new mortgage loan is opened immediately following the loan's last report, the loan's borrowers

move, or the loan was very near the end of its term, the loan is assumed to have closed at the date of last report. In the remaining cases, the loan is treated as provisionally open for two years. If at that point no update has been received (as should be the case with a servicing transfer) the loan is treated as closed.

Experian data are also used to construct monthly values of loan performance, balance and payments. Construction of loan performance is discussed separately in Appendix B. Unlike performance (in which the final dataset contains missing values and nothing is imputed), it was decided to include in the final dataset complete loan balances and required payments for every month for every loan from its origination (or January 1998) to termination.

Historical monthly data on balance plus required and actual payments (going back 24 months) are included in the core credit bureau files and thus subject to the dispute process which can lead to cleaner data. They have been included only since 2010 though and are not present for every loan. Moreover, unlike performance, historical balance and performance data are constructed and maintained by Experian and not reported each month by servicers. NMDB data were supplied by Experian semi-annually prior to 2012 and quarterly until 2020 when the NMDB moved to a monthly schedule. Consequently, in constructing complete monthly vectors of balance and payments it is necessary to deal with significant gaps, particularly for loans originated prior to 2010.

When possible, monthly balance gaps were dealt with by interpolating from the two end points using the inferred appropriate amortization rate. This process works most of the time, but breaks down when loans are in dispute, forbearance, or default or buyers make extra or non-standard payments (*e.g.*, curtailment). Here it was necessary to use linear or other forms of interpolation. Loans missing balance information in the beginning were generally assumed to have prepaid interest. Non-delinquent loans that had missing data at the end point were extrapolated using inferred amortization rates. In a few rare cases, the entire monthly balance vector had to be constructed using assumptions about amortization.

Only the required payment vector was filled; not the actual payment which is left unedited. Here, it was generally assumed that missing required payments were the same as the last known value. For ARMs with required payments changing once a year, that information is taken into account in constructing the infill. Again, in a few cases the entire payment vector needed to be constructed. Here, contract interest rate and balance information were used to infer the likely required payment.

The remaining editing and cleaning dealt with variables added with the matching process. Most prominent of these is loan purpose (*e.g.*, home purchase, refinance) which is not directly available in the Experian data. It is, however provided in almost all cases when there is an administrative or HMDA match. In about 9% of these cases, though, HMDA reporting is inconsistent with the administrative data and even within the same source, purpose data is inconsistent almost 1% of the time. This is a particular problem with loans which do not fall neatly into either the purchase category or the refinance category, such as when a borrower buys out their co-borrowers. To deal with this problem, the 11% of overall loans which involve a change in borrowers but not property are classified as a separate purpose category (change of

borrowers) to let users decide how to treat them.

Loans with missing purpose are coded by examining the borrowers' other loans and potential moves. If there is no preceding loan or the timing of the termination of the previous loan is inconsistent with a refinance, the loan is classified as home purchase. When the timing is consistent with both a purchase or refinance, it is treated as a purchase if all borrowers moved at the time of origination, a refinance when there is no change of borrowers and at least one borrower did not move, and a "change of borrowers" in all other cases. Finally, in about 3 percent of the cases, a loan was taken out so long ago (prior to July 1991) that no information on the previous loan is available in the credit files. In these instances, the loan purpose is imputed using statistical methods with the outcome variable constrained to be home purchase or refinance.

Property location is also not supplied directly in the Experian data. The final dataset includes the state, county and census tract that the property associated with the loan is estimated to be in 2019 and in the year the loan was originated. For loans with an administrative match, the address supplied by the administrative partner is geocoded by Experian and tract information supplied to the NMDB. For other loans, Experian maintains addresses for each person in their files compiled from the billing addresses supplied by loan servicers augmented by addresses from marketing sources. Going back to archival files, Experian was able to identify and supply to the NMDB team information on up to 57 addresses associated with each borrower. For each address, Experian provided the ZIP Code and 2000 and 2010 census tracts associated with the address, as well as the date the address was first linked to the borrower.

Tract assignments for mortgages not matched to administrative files is determined from these data. For example, for all active mortgages, the tract of the most recently reported address is used. For closed loans, addresses are given preference when the borrower moves (and the new address is reported to Experian) around the time the loan was originated. Respondents participating in the NSMO or ASMB are asked if the property is located at the address used in the mailing. If so, then that address is prioritized in assigning the tract; if not, then it is not allowed to be assigned.

Assignment of both the original and 2019 tract was straightforward for almost all NMDB loans originated in 2003 or later because Experian provided both 2000 and 2010 tracts for each address in the database. From 1992 to 2002, however, the Census relied on the 1990 Census tract taxonomy. Prior to 1992 the 1980 census taxonomy applied which left out much of rural America which was not tracted. Assigning origination year tracts to NMDB loans originated during these periods represented a challenge. Administrative data and tract assignments available from HMDA provided 1990 or 1980 origination-year tract numbers for some of these older NMDB loans. However, for most others, it was necessary to first identify the 2000 census tracts associated with the property backing the loan from information available in the Experian address file and then use Census Bureau crosswalk files to assign the 1990 tract and a NMDB-constructed crosswalk based tract centroids to assign 1980 tracts. In a small number of cases, only the 1980 or 1990 tract was available, so the crosswalk files needed to be used to determine the appropriate 2019 tract to assign to the loan.

Another variable requiring some editing/cleaning is loan type (FHA, VA, or RHS) and GSE type (Fannie Mae, Freddie Mac, or FHLB). For the most part, both variables are determined by the administrative loan matching process. Loans confirmed as matches to VA administrative data are designated as VA loans, for example. These designations do need to be supplemented for loans newly added to the database, which have not yet gone through the matching process, and for older loans, where the matching process is less accurate. In these cases, Experian "account-type" codes are used to identify FHA, VA, and RHS loans and "secondary loan type" codes are used to identify Fannie Mae and Freddie Mac loans. For older loans, identifiers derived from HMDA data matches are also used. It should be noted that loan type is a characteristic of the loan and does not change during its lifetime. GSE acquisition though is time dependent. NMDB flags a loan as GSE if it is ever owned by a GSE. This means that the classification of a loan can change in the NMDB if it is not sold to the GSEs until it is "seasoned."

The majority of the remaining editing and cleaning is devoted to implementing rules to determine values for four key borrower variables—age, gender, race, and ethnicity—and nine key loan variables—borrower income, property value, LTV/CLTV, contract interest rate, debt-to-income ratio, occupancy status (owner-occupied or not), property type (manufactured housing or not), number of units in the property (1 to 4), and ARM status.

Values for most of these variables are derived from administrative or HMDA data matches. In most instances, determining the variable value is straightforward. In some cases, however, administrative sources may have conflicting values, or the values may be implausible. Sometimes these issues are obvious and easy to resolve— for example, when LTVs are reported in ratios instead of percentages, or when monthly instead of annual income is reported. In other cases, the best resolution is less clear-cut. In general, the waterfalls developed for the NMDB to resolve these data issues prioritize plausible values over implausible and data collected from high quality administrative matches (those done by Experian with data partners and matched by SSN) over data from fuzzy logic matches (those done with HMDA and McDash data). These rules also consider internal consistency. For example, an administrative data match implying both a very high borrower income and a very high debt-to-income ratio might be given a low weight. Contract interest rates which are inconsistent with the implied amortization rate derived from balance patterns also might be deprioritized or disallowed.

Because borrower characteristics are not loan dependent, values of these variables are dealt with in a different way. NMDB includes birth month and year, gender, race, and ethnicity for every NMDB borrower. Birth month and year are populated for almost all NMDB borrowers because they are carried and collected as part of a credit report. Gender is not included as a credit variable but is widely and accurately available through Experian's marketing database. Gender is also available from HMDA data, as is borrower age beginning with the 2018 data. Both variables are also available for survey respondents and, in some cases, from administrative data. When consistent with the timing of mortgage loan originations (for example, borrower need to be at least 18 when their earliest mortgage loan was taken out), the Experian-based values were used. Otherwise, when values for age or gender were inconsistent or, in the case of age, implausible, they were imputed (see Appendix D) rather than assigned through waterfall priorities.

High quality data on race and ethnicity are not available from Experian. Fortunately, such information is available from HMDA data and from administrative sources. Moreover, since all of an NMDB borrower's loans, sample or not, are potentially matched to both HMDA data and the primary administrative data, only one of their loans needs to have matched and have race and ethnicity data populated. The best and least ambiguous case is when the matched loan has a single borrower. However, even with multiple borrowers, race and ethnicity are most often reported as the same for all, so the assignment is still straightforward. In other cases, the borrower's gender can be used to determine which HMDA data field or administrative data field to use. In the relatively small number of cases where external matches provide inconsistent or ambiguous values, the values are imputed rather than subjected to a waterfall.

One issue that arises in determining race and ethnicity is the change in collection rules for HMDA data beginning in 2004. Prior to that date, HMDA data treated ethnicity as a race, effectively forcing reporters to report either race or ethnicity, but not both. Moreover, reporters were allowed only one racial classification (that is still the case with some administrative data). In these instances, the variable not reported (*e.g.*, race if the HMDA variable were reported as "Hispanic") needs to be imputed (see Appendix D). Ultimately, NMDB provides age, gender, race and ethnicity for each borrower associated with every sample mortgage.

Appendix D. Imputation of Missing Data in the NMDB

Early in the development of the NMDB, it was decided that missing values of key variables would be imputed in the NMDB. This is a common practice for many statistical databases, including, for example, most census products. Obviously, this makes the dataset easier to use, but more importantly it also likely makes the statistical analysis using the dataset more accurate. Unless missing data occurs completely at random, analysis using loans only with no missing data is likely to produce bias. Statistical imputation, as used in the NMDB and described in this Appendix, requires the weaker assumption that missing data occurs randomly conditioned on the predictor variables used in the imputations. Recognizing that some analysts may prefer to deal with missing information themselves, indicator variables are included in the NMDB which identify which specific variable values were imputed.

In Appendix C above, methods of dealing with missing information, such as monthly information on loan balances and payments, through interpolation and editing procedures are addressed. Here, methods of dealing with missing values which cannot be addressed that way are presented.

Five continuous variables (borrower income, property value, LTV/CLTV, contract interest rate, and debt-to-income ratio) and nine categorical variables (borrower age, gender, ethnicity, and race, and mortgage purpose, occupancy status (owner-occupied or not), property type (manufactured housing or not), number of units in the property (1 to 4), and ARM indicator (yes if ARM, no otherwise) are subject to imputation. Different methods are used for each variable.

Borrower birth month and year are populated for almost all NMDB borrowers because they are carried and collected as part of a credit report. Gender is not included as a credit variable but is widely available through Experian's marketing database. Both variables are partially available from some of the administrative sources as well as HMDA data. Nevertheless, values need to be imputed in a relatively small number of cases. "Hot deck" methods are used for these imputations. Borrowers are matched to cells of borrowers with complete gender and age information based on the relative age and gender of co-borrowers, loan size, age of oldest tradeline, and date of origination. A random match within the cell is selected to assign values. Age and gender are assigned iteratively: age first, then gender, then age is re-estimated based on the assigned gender.

Values for the remaining 12 variables are assigned using logistic or linear regression models. The imputation models are constructed from NMDB data as well as demographic (tract or county based) data from HMDA data and other sources. The tract-level data, such as the median borrower income level in HMDA data of all borrowers in that tract in the same year the NMDB loan was originated, in particular, are powerful predictors. The models are defined by time period and mortgage characteristic group. The intent of the models is to provide the statistically best unbiased estimated value of the imputed variable.

Mortgage purpose (*i.e.*, home purchase or refinance) is imputed separately from the other variables. It is imputed only for older loans (originated pre-1992) when HMDA data are not available. The property type variable is available in the NMDB beginning with 2004

originations, thus is not included in the early year models. Otherwise, the imputation estimation is performed recursively in a specific order, such that the assigned imputed values from earlier models in the scheme can be used as predictor variables in models later in the scheme. Of the categorical variables, ethnicity and race are imputed on the borrower level data. The rest of the variables are imputed on the loan level.

Imputation models are based on different time periods which is a tradeoff between flexibility of the modelling process and the availability of explanatory variables. Starting in 1992, all loan level imputation models are estimated and applied separately for each year based on the origination year of the loan through 2019. Loans originated prior to 1992 are imputed using single models for each variable with "time effects" as intercept shifts rather than completely different models. Values for loans originated in 2020 are based on estimates from the 2019 models. When new HMDA data are available for 2020, a separate model will be created for that year and used to impute 2021 as well until 2021 HMDA data are available.

With each quarterly update of the NMDB, the imputation models for the most recent three origination years are re-estimated. This allows the size of the dataset used to estimate the models, which rely on loans with "hard data," to expand naturally as more NMDB loans are matched to administrative loan files with such data. Similarly, missing values of NMDB loans originated in these three years are re-imputed each quarter to capitalize on the improved models. Otherwise, the predictive models are "frozen" but still can be used to impute values for missing variables with loans newly added to the NMDB.

The two borrower level imputations are estimated based on three models each. These models reflect different periods defined by the availability of HMDA data. The periods are: 2004 to present, 1992 to 2003 and pre-1992. Prior to 2004 originators were able to report either race or ethnicity, but not both, in HMDA data. Thus, if only pre-2004 HMDA data were available for a borrower (many borrowers have multiple HMDA data matches in the NMDB) special imputation models needed to be used to predict ethnicity given race or vice-versa. These models supplemented models used when all information was missing.

Imputation follows a generic process for each variable. First, models are estimated using loans or borrowers with complete information. Then, values for loans or borrowers with missing information are assigned. For continuous variables, the assignment is based on the predicted "y-hat" from a standard regression model plus a random error term drawn from the actual error distribution divided into deciles based on the values of "y-hat" (thus the error assigned to a loan with a predicted "y-hat" in the top decile would be restricted to actual errors of loans used in the model estimation who also had "y-hats" in the top decile). Final assignments are bounded by the 1st- and 99th-percentiles of the distribution for known values. Binary and multi-nominal logistic models are used to impute categorical variables with specific assignments based on random draws conditioned on the probabilities assigned by the logistic models.

Once estimates for missing values were developed, the overall distribution, and for various subsets, of the estimated values were compared to the overall and corresponding subset distributions of the existing values to ensure that they are consistent. The selection of the specific regressors and model forms was based on econometric goodness of fit tests and the

application of various robustness checks.

The imputation process follows the following recursive order: borrower income, property value, LTV/CLTV, contract interest rate, and debt-to-income (all continuous). Then occupancy status and property type (both dichotomous), number of units (1 to 4), and ARM indicator. For borrowers, ethnicity is estimated first (dichotomous) and race second (multi-nominal).

The percentage of each variable that is imputed in the NMDB is given in table D-1. Statistics are presented separately for three periods defined by different HMDA data rules and for loans associated with NMDB administrative data partners (Fannie Mae, Freddie Mac, FHLB, FHA, VA, and RHS) and those that are not. Data for 2020 loans are not shown as "hard data" since matching for them is not complete yet.

Table D-1. Percent Imputed

	Before 2004		2004 to 2011		2012 to 2019	
Loan Level	Data Partner	Other	Data Partner	Other	Data Partner	Other
Borrower Income	14.3	44.5	5.6	42.8	7.0	52.8
Property Value	13.1	78.3	3.1	46.4	4.5	72.0
LTV/CLTV	13.1	78.3	3.1	46.4	4.5	72.0
Contract Interest Rate	6.9	36.4	1.3	24.8	2.7	32.7
Debt-to-Income Ratio	23.7	94.6	7.7	78.1	11.3	86.2
Occupancy Status	7.2	37.1	1.4	24.5	3.7	48.8
Property Type	-	-	12.3	40.1	32.8	58.8
Number of Units	10.8	81.2	3.1	51.5	6.6	74.2
ARM Indicator	44.2	78.0	21.0	46.1	18.3	72.7
Borrower Level	All		All		All	
Birth month	0.2		0.1		0.1	
Gender	2.7		1.4		1.9	
Race	14.7		8.9		16.5	
Ethnicity	42.6		12.3		16.7	

The specific models and sample groupings used in this process and summary statistics from the estimation are given in tables D-2 and D-3.

Table D-2. Imputation Models, 1992 - Present				
			Adjusted R^2 Range	
Variable	Allowable Values	Estimation Model Grouping	Min	Max
Borrower Income	Continuous	Purchase, FTHB*	37.7%	63.7%
		Purchase, Non-FTHB	38.8%	50.5%
		Refinance	37.0%	52.7%
Property Value	Continuous	Purchase, FTHB	75.8%	94.4%
		Purchase, Non-FTHB	77.8%	90.6%
		Refinance	74.0%	87.8%
LTV / CLTV	Continuous	Conventional Purchase, FTHB	36.2%	63.5%
		Conventional Purchase, Non-FTHB	35.6%	55.7%
		Conventional Refinance	35.3%	55.2%
		Government Purchase, FTHB	6.6%	46.9%
		Government Purchase, Non-FTHB	12.7%	68.0%
		Government Refinance	22.6%	64.5%
Contract Interest Rate	Continuous	Purchase, FTHB	9.8%	44.1%
		Purchase, Non-FTHB	11.6%	39.6%
		Refinance	9.6%	23.4%
Debt-to-Income Ratio	Continuous	Purchase, FTHB	34.2%	48.4%
		Purchase, Non-FTHB	27.2%	38.0%
		Refinance	28.9%	48.8%
Ethnicity	0 = Non-Hispanic	Purchase, FTHB	57.7%	67.4%
	1 = Hispanic	Purchase, Non-FTHB	49.6%	58.8%
		Refinance	47.4%	58.6%
Race	1 = White	Not Applicable	NA	NA
	2 = Black		86.3%	89.8%
	3 = American Indian		43.8%	47.1%
	4 = Asian		81.8%	88.4%
	5 = Hawaiian		48.0%	57.5%
	6 = More than one race, Other Black		48.3%	67.5%
	7 = More than one race, Other Non-Black		53.0%	57.2%
Occupancy Status	0 = Not Owner Occupied	Purchase	39.7%	42.2%
	1 = Owner Occupied	Refinance	29.4%	31.3%
Property Type	1 = Site Built	Purchase	51.2%	54.2%
	2 = Manufactured Housing	Refinance	37.6%	39.4%
Number of Units	1 - 4 Units	Owner Occupied	29.5%	31.3%
		Not Owner Occupied	24.1%	25.7%
ARM Indicator	0 = Not ARM	Purchase	17.1%	18.5%
	1 = ARM	Refinance	19.2%	25.7%

Note: *FTHB is first-time homebuyer

Table D-3. Imputation Models												
	Loan Purpose	Borrower Income	Property Value	LTV / CLTV	Contract Interest Rate	Debt-to-Income Ratio	Ethnicity	Race	Occupancy Status	Property Type	Number of Units	ARM Indicator
Model Type	Logistic	OLS (Log-Log)	OLS (Log-Log)	OLS	OLS (Log-Log)	OLS (Log-Log)	Logistic	Multi-Nomial Logistic	Logistic	Logistic	Logistic	Logistic
Explanatory Variables												
Loan Purpose - Purchase/Refinance											X(c)	
Loan Amount	X	X	X	X	X	X			X	X	X	X
Borrower Income			X		X	X	X	X	X	X	X	X
Income Ratio				X								
Property Value				X								
LTV					X	X			X	X	X	X
Rate Spread						X			X	X	X	X
DTI									X	X	X	X
Ethnicity								X(c)				
Occupancy Status										X(c)		X(c)
Primary Market Interest Rate					X	X			X	X	X	X
Credit Score at Origination		X	X	X	X	X	X	X	X	X		X
Number/Gender of Borrower(s)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)			X(a)			
Number of Borrowers								X(c)				X(c)
Gender of Borrower							X(c)	X(c)				
Age of Borrower	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)		X(a)
Hispanic Borrower (marketing data)							X(c)					
Hispanic Borrower (country of origin)							X(c)	X(c)				
Minority Borrower		X(c)	X(c)		X(c)	X(c)		X(c)				
Ethnic Category based on Surname (marketing data)								X(a)				
Education Level of Borrower	X(a)	X(a)					X(a)	X(a)	X(a)	X(a)		X(a)
Tract Median Income		X	X	X	X	X						
Tract Median Loan Amount		X	X	X	X	X						
Tract Median Property Value		X(b)	X	X	X	X			X(b)	X(b)	X(b)	X(b)

Table D-3. Imputation Models (Cont.)												
	Loan Purpose	Borrower Income	Property Value	LTV / CLTV	Contract Interest Rate	Debt-to-Income Ratio	Ethnicity	Race	Occupancy Status	Property Type	Number of Units	ARM Indicator
Model Type	Logistic	OLS (Log-Log)	OLS (Log-Log)	OLS	OLS (Log-Log)	OLS (Log-Log)	Logistic	Multi-Nomial Logistic	Logistic	Logistic	Logistic	Logistic
Explanatory Variables												
Tract Median Debt-to-Income Ratio		X	X	X	X	X						
Tract Median Monthly Loan Payment		X										
Tract Median LTV			X	X	X	X						
Tract Percent Owner-Occupied									X(d)	X(d)	X(d)	X(d)
Tract Percent Refinance Originations	X											
Tract Percent Minority		X	X		X	X						
Tract Percent Racial Category								X(d)				
Tract Percent Hispanic							X	X				
Ratio: Tract Median Loan Amount to Median Property value		X										
Loan Made at Beginning of Month		X(c)			X(c)	X(c)						
Loan Made at End of Month			X(c)		X(c)	X(c)						
Non-Conforming Loan		X(c)	X(c)		X(c)	X(c)						
FHA Eligible Loan		X(c)	X(c)	X(c)	X(c)	X(c)						
Borrower has taken out VA Loan Before	X(c)	X(c)	X(c)	X(c)	X(c)	X(c)						
Small Loan	X(c)	X(c)	X(c)	X(c)	X(c)	X(c)						
Tract Median income associated with Small Loans		X	X									
Tract Median Loan Amount of Small Loans			X									
Tract Median Debt-to-Income Ratio of Small Loans						X						

Table D-3. Imputation Models (Cont.)												
	Loan Purpose	Borrower Income	Property Value	LTV / CLTV	Contract Interest Rate	Debt-to-Income Ratio	Ethnicity	Race	Occupancy Status	Property Type	Number of Units	ARM Indicator
Model Type	Logistic	OLS (Log-Log)	OLS (Log-Log)	OLS	OLS (Log-Log)	OLS (Log-Log)	Logistic	Multi-Nomial Logistic	Logistic	Logistic	Logistic	Logistic
Explanatory Variables												
Year of Origination	X(a)											
Origination Month of the Year	X(a)		X(a)	X(a)	X(a)							
Location - State		X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)	X(a)
Location - Metro Area								X(c)	X(c)	X(c)	X(c)	X(c)
Location - Rural Area									X(c)	X(c)		
Term of Loan	X			X	X	X						X
Short-Term Loan	X(c)			X(c)	X(c)	X(c)						
Loan Closed During current year					X							
FHA, VA, RHS Insured/Guaranteed	X(c)											
FNM, FRE, or FHLB Loan									X(c)			

Note: (a) Multi-Categorical Binary Variables (b) Beginning in 2018 (c) Binary Variable (d) Tract share by multiple categories
(Words in parentheses) indicate where that data came from/how it was derived


BUREAU OF CONSUMER FINANCIAL PROTECTION
UNITED STATES OF AMERICA


FEDERAL HOUSING
FHFA
FINANCE AGENCY