The National Mortgage Database (NMDB)

Robert Avery, Ken Brevoort, Theresa DiVenti, Carla Inclan, Ian Keith, Jessica Lee, Lexian Liu, Ismail Mohamed, Forrest Pafenberg, Jay Schultz, Cynthia Waldron, Xun Wang, Claudia Wood, Peter Zorn

Federal Housing Finance Agency
Consumer Financial Protection Bureau
Freddie Mac

Association of Public Data Users Annual Conference
September 16, 2013

The views expressed are those of the authors and do not necessarily represent those of the Consumer Financial Protection Bureau, the Federal Housing Finance Agency, Freddie Mac or their staff.
What is the NMDB?

- A new, *nationally representative*, loan-level mortgage database jointly funded and managed by the FHFA and CFPB based on a prototype developed by Freddie Mac.
  - 1st lien mortgages reported to the credit bureaus are used as both the sampling frame and the source of performance data.
  - No new data is collected—the NMDB will make better use of data that already exists. No personally identifiable data in NMDB dataset. Serious effort to address privacy issue.
  - The database is a 1/20 sample. Has the advantage of much less cleaning than a universe or registry, much better privacy protection, and almost the same statistical precision.
  - Because the credit bureaus archive their data, the NMDB recovers data that would have been available had the project been started years ago.
  - The initial 1/20 sample is representative of all mortgages open at any time from January 1998 to June 2012 and any borrower who had at least one mortgage during that period.
  - Going forward, a 1/20 representative sample of newly originated mortgages will be added each quarter, and terminated mortgages will exit the sample.

- Credit bureau data are comprehensive. However, they are raw servicing data which requires significant cleaning and matching to additional data sources to make them useful.
  - Much of the cleaning is being done by FHFA staff.
  - Working with active cooperation of credit bureau (Experian) staff.

- NMDB will also have a survey component. Each quarter a representative subset of borrowers associated with loans newly added to the database will be sent a mail survey soliciting information on their mortgage shopping and origination experience.
Four Overlapping Databases

- The unit of observation for the primary database is the mortgage.
  - There are 10.1 million sample mortgages in the initial mortgage database.
  - The mortgage database will contain full credit information for all borrowers (currently 14.7 million) associated with the sampled mortgages.
  - Borrower data will be gathered from one year prior to sampled mortgage origination to one year after termination and tracked quarterly.
  - Performance on the sample mortgages will be collected monthly.

- The NMDB will also make available an historic database containing full credit data (including scores) from 1998 to 2012 of borrowers associated with any active sample mortgage during the 1998 to 2012 period.
  - The database will contain information on all mortgages (47 million) taken out by these borrowers during the 1998 to 2012 period. This represents almost 25% of the mortgages active at some point during this period.
  - Performance for each mortgage and all other consumer obligations will be tracked from 2000 to 2012.

- The NMDB will maintain a separate database of a representative 1-in-20 sample of individuals who have ever had an active mortgage from 1998 onward (at present about 6 million persons).
  - Quarterly information will be maintained on these individuals from one-year prior to taking out their first mortgage (or 1998) until they die or are removed from the credit bureau.
  - Persons will be added to the database when they take out their first mortgage.

- The NMDB origination survey data will be maintained as a separate database.
Why Does the Data Have to Be Cleaned?

- Credit Bureau Data are collections of “tradelines” (monthly reports of account performance from one servicer). Many mortgages have a sale of servicing or transfers within an organization which leads to multiple tradelines for one mortgage. They have to be “deduped.”

- Tradelines are purged from the active system 10 years after they are closed. Derogatory performance information is purged after 7 years. Historical archives have to be recovered to add purged information back into the database to create a complete historical database.

- The Credit Bureaus assign each tradeline to a “person.” Many times these assignments need to be changed as they learn that John Q Smith is also J. Smith. The entire history of such reassignments needs to be recovered if old archival data are to be merged into the current system.
Why does Data Have to be Added?

- Raw Credit Bureau data lack some key items need to assess the underwriting risk of a loan and to understand the circumstances of the borrowers:
  - Property characteristics and value (e.g. purchase price, appraised value),
  - Borrower characteristics (e.g. age, race, income, gender),
  - Some mortgage characteristics (e.g. ARM status, PMI, origination channel).

- The database is being supplemented with information obtained from matching to existing external sources (some still under negotiation). These include:
  - Administrative files (FHA, VA, Fannie Mae and Freddie Mac). We estimate that about 68% of sample loans originated between 1998 and 2012 are in these files. High quality matching using PII (borrower name, address) with 3rd party blind matching process should yield almost all needed mortgage and property characteristics for these loans.
  - Third party servicing data (e.g. LP, LPS) and Private label MBS data. We estimate that about 14% of sample loans originated between 1998 and 2012 are in Private label MBS loan-level data files, and thus matchable to the NMDB using property address. About 10 percent more loans are estimated to be in available third party servicing databases.
  - Property transaction, MLS and appraisal datasets matched by borrower address can provide property characteristics for many of the remaining loans
  - Home Mortgage Disclosure Act (HMDA). We have obtained an 80% high quality match rate using fuzzy logic based on property census tract, loan origination date and amount. A HMDA match gives income and race.
  - Household moving/address information on up to 25 last addresses and data on age, gender and marital status are available from public records collected by the credit bureau.

- The challenges will be the 20% of sample loans originated before 1998 (including 4 percent before 1990); the estimated 9% that are non-owner-occupied; and 10% that are under $50,000. Each of these loan types are more difficult to match to external records.
Why Can’t Existing Databases be Used?

- **HMDA:**
  - Pretty close to a universe of all 1st lien mortgage originations.
  - However, has limited borrower, property and loan information.
  - Critical flaw is that it has no performance data.

- **LPS McDash and/or CoreLogic:**
  - Servicing files from 25-30 large servicers versus 2,000 servicers in credit bureaus.
  - Not representative—no coverage of smaller servicers.
  - Same problems as underlying NMDB data—duplication, hanging performance and sale of servicing—but not cleaned as the NMDB will be, so you don’t know it.
  - No information on other obligations—particularly 2nd liens--previous or subsequent mortgages,. Little information on borrowers.

- **NY Fed Equifax:**
  - Similar source as NMDB, but unit of observation is borrowers not loans.
  - Same problems as underlying NMDB—duplication, hanging performance and servicing sales—but not cleaned as the NMDB will be, so you don’t know it.
  - Little or no supplementation with other data. Difficult to link files over time.
Timeline

- Contract signed with Experian on September 27, 2012. FHFA NMDB Team assembled in the Fall of 2012.
- An analytic group at FHFA, Freddie Mac and CFPB is processing and cleaning the data and will match it to external sources, impute data for loans that cannot be matched, and develop a series of regular reports and queries to facilitate use of the NMDB.
  - It will likely take until spring or summer 2014 to finish cleaning the data.
  - Some data items will be available by the beginning of 2014.
  - Property value (needed for LTV) will likely take the longest to add.
  - Origination survey will be initiated in October 2013. Almost through OMB.
- An existing pilot prototype dataset in development for 2 ½ years funded by Freddie Mac (1/500 sample of loans outstanding since 2003) is available.
  - Prototype will be maintained and was just updated in May 2013.
  - Already used in FHFA’s 2012 HERA-mandated report.
  - The Prototype was used to pilot test the Origination Survey and (currently) a Delinquency Survey. A 60 percent response rate was achieved on the Pilot.
Access and the Future

- NMDB is being set up as a public good. We believe that the contract signed with Experian is a model for such data.
- The challenge is to (1) protect borrower/lender personally identifiable information and (2) provide users access to useful data. Key points of conflict on these objectives are items such as: property location, loan amount and timing of origination which are critical for mortgage analysis but potentially could be “reverse-engineered” to identify specific borrowers.
- Our solution:
  - NMDB loan-level data is physically housed *only* on a FHFA/CFPB server. No PII or geography below census tract is contained in the dataset. No other data containing PII are allowed on the server.
  - Access to the loan-level data is allowed for any federal government/Reserve Bank/Fannie Mae/Freddie Mac employee going through the following access process:
    » Must sign an agreement not to reverse engineer identity of borrower or lender. Severe penalties for violations of agreement.
    » All work behind a firewall—data can’t be removed. Users will be able to request that non-PII data be added to the servicer but follow strict protocalls to do so.
    » Data cannot be used for “enforcement.” Users will not have the identity of the servicer.
  - Access to aggregated NMDB data not representing privacy risk will be available to the general public through something like “American Fact Finder.”
The Data so far: Total Volume (1st lien loans over $50k)
The Data so far: Overall Coverage is Good
The Data So far: Can now look at Share
Example of Analysis – Data from the Prototype
Default rates are higher for firsts with seconds

Default (90d or worse)

First with Second
First without Second

First Lien Open Date

First with Concurrent non HELOC
First with Concurrent HELOC
First with Subsequent non HELOC
First with Subsequent HELOC
First without Second
Example of Analysis – Data from the Prototype
Performance of firsts w/ different types of seconds

<table>
<thead>
<tr>
<th>Seconds and Firsts perform similarly</th>
<th>Concurrent ALL</th>
<th>HELOC</th>
<th>non HELOC</th>
<th>Subsequent HELOC</th>
<th>non HELOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seconds and Firsts perform similarly</td>
<td>87%</td>
<td>89%</td>
<td>87%</td>
<td>88%</td>
<td>78%</td>
</tr>
<tr>
<td>Seconds perform better</td>
<td>8%</td>
<td>7%</td>
<td>8%</td>
<td>7%</td>
<td>15%</td>
</tr>
<tr>
<td>Seconds perform worse</td>
<td>5%</td>
<td>4%</td>
<td>5%</td>
<td>4%</td>
<td>7%</td>
</tr>
</tbody>
</table>

- 88% of firsts and their associated seconds perform similarly
- When performance diverges, seconds tend to out-perform their associated firsts.
- GSE firsts and their associated seconds perform better than non-GSE loans.
- Firsts with piggyback non-HELOC (closed end) seconds have the highest default rates.
Originations Survey—Sample Results from the Pilot

How well could you explain to someone...

- Difference between a fixed and adjustable rate mortgage: 69% Very, 23% Somewhat, 3% Not at all
- Repercussions of not making required mortgage payments (default): 58% Very, 31% Somewhat, 11% Not at all
- Process of taking out a mortgage: 45% Very, 51% Somewhat, 5% Not at all
- Difference between a mortgage interest rate and APR: 30% Very, 45% Somewhat, 25% Not at all
- Difference between a prime and subprime loan: 27% Very, 37% Somewhat, 35% Not at all
- Mortgage disclosure forms: 17% Very, 53% Somewhat, 31% Not at all

In the process of getting your mortgage did you...

- Total: Answer request for more information 60%, Resolve credit errors or problems 16%, Delay, postpone closing date 13%, Change lenders 13%, Add co-signer 5%, Multiple appraisals 8%
- New Loan Borrower: Answer request for more information 73%, Resolve credit errors or problems 21%, Delay, postpone closing date 20%, Change lenders 12%, Add co-signer 6%, Multiple appraisals 11%
- Refinancer: Answer request for more information 51%, Resolve credit errors or problems 10%, Delay, postpone closing date 9%, Change lenders 14%, Add co-signer 5%, Multiple appraisals 6%

National Mortgage Database