Mortgage and Housing Microdata: Database Construction and Challenges for Statistical Inference

Nancy Wallace

Haas School of Business
Real Estate and Financial Markets Laboratory
Fisher Center for Real Estate and Urban Economics

Macro Financial Modeling Conference
January 29, 2016
Statistical Inference in the Mortgage Market

► Standards should be very high given policy objectives.

► Mortgage origination is a high dimensional contracting problem so most contract elements are jointly endogenous.

► Lenders set the contract menus based on pricing embedded options (default and prepayment) – leads to *ex post* sample selection and sorting in mortgage origination data.

► Pre-crisis mortgage regulation controlled by competing regulatory authorities, Federal TILA, state consumer protection and foreclosure laws, zip code level Underserved Area Goals.
  ● Induces non-random assignment and rationing of contract types even at the level of zip codes.

► No standardized loan identification number exists in U.S. – leads to challenges in data set assembly and sampling.

► Available mortgage data sets are problematic because none of them span the contracting space.
No Single Data Set Span the Contracting Space

<table>
<thead>
<tr>
<th></th>
<th>HMDA</th>
<th>LPS/Equifax</th>
<th>ABSNet/Lewtan</th>
<th>DataQuick/Corelogic</th>
<th>FNMA/FHLMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>~90% Orig.</td>
<td>~20% Orig.</td>
<td>~100% PLS</td>
<td>Liens and Transfers</td>
<td>N.A.</td>
</tr>
<tr>
<td>Storage (G)</td>
<td>100</td>
<td>8,500</td>
<td>360</td>
<td>2,100</td>
<td>130</td>
</tr>
<tr>
<td>Loans (M)</td>
<td>262.97</td>
<td>67.40</td>
<td>20.00</td>
<td>118.47</td>
<td>31.31</td>
</tr>
<tr>
<td>Orig. ($ T)</td>
<td>44.70</td>
<td>11.50</td>
<td>4.48</td>
<td>36.48</td>
<td>5.61</td>
</tr>
</tbody>
</table>

**Static Fields Loan Level**

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Primarily GSE</th>
<th>Non-GSE</th>
<th>All</th>
<th>GSE FRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Types</td>
<td>All</td>
<td>Primarily GSE</td>
<td>Non-GSE</td>
<td>All</td>
<td>GSE FRM</td>
</tr>
<tr>
<td>Geographic Identifiers</td>
<td>Census Tracts</td>
<td>Zip Codes, Zip (3) GSE</td>
<td>Zip Codes</td>
<td>Address</td>
<td>Zip (3)</td>
</tr>
<tr>
<td>Borrower Income</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>DTI</td>
</tr>
<tr>
<td>Interest Rate</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Orig. Date</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Orig. Amount</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Originator</td>
<td>Yes</td>
<td>No</td>
<td>Aggregator</td>
<td>Yes</td>
<td>Aggregator</td>
</tr>
<tr>
<td>House Price</td>
<td>No</td>
<td>Yes – LTV</td>
<td>Yes – LTV</td>
<td>Yes</td>
<td>Yes - LTV</td>
</tr>
<tr>
<td>Other Liens</td>
<td>No</td>
<td>If Equifax</td>
<td>Partial</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Sec. Status</td>
<td>Partial</td>
<td>Yes</td>
<td>No</td>
<td>Partial</td>
<td></td>
</tr>
</tbody>
</table>

**Dynamic Fields Loan Level**

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>YMD</th>
<th>YMD</th>
<th>Payoff YMD</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payment Performance</td>
<td>No</td>
<td>YMD</td>
<td>YMD</td>
<td>Payoff YMD</td>
<td>Y</td>
</tr>
<tr>
<td>House Price</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes, not attributes</td>
<td>No</td>
</tr>
<tr>
<td>Credit Score</td>
<td>No</td>
<td>If Equifax</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Linking mortgage data: Without Loan IDs

Data

Technology

Analytics

High Performance Data Management: Cloudera/Impala/Hadoop
High Performance Computing Cluster(HPCC)

Mortgage Population

HMDA  ABSNet Mort.  GSEs  LPS

Liens  ABSNet RMBS  EDGAR Bond CUSIP  Other Credit

RMBS Bonds

Housing Market Characteristics  DataQuick Transactions

Micro Network Model Competition & Funding

Micro Loan Level Stress Test

Micro Corporate Stress Test

Macro Stress Test

Flow of Funds  Y-9C

Equifax

EDGAR

Corporate

Stress Test

Interest Rate Model

House Price Model

Micro

Loan

Population

Competition & Funding

High Performance Data Management: Cloudera/Impala/Hadoop
High Performance Computing Cluster(HPCC)
Average match rates in the literature are usually around 30% or less.

DataQuick and ABSNet Merge (2000 - 2013): For zip codes that are present in DQ (90% coverage in U.S.) overall match is 87%, 97% for California.

DataQuick and HMDA Merge (2005 - 2013): Overall match of 60%, 90% for California.

LPS and DQ merge is currently underway.

Exploit network structure of lenders and subsidiaries to enhance matches.
Why does this Matter?

- Challenges of sampling and population inference using existing data sets.
- Example 1: Inference concerning the competitive structure of the mortgage market.
- Example 2: Inference concerning house price dynamics and mortgage performance.
Herfindahls using only HMDA: Market Appears Competitive

Stanton et al. (2014)
Mortgage Market Loan-Flow Networks

1.0 Bank/Thrift Holding Company
  1.1 Holding Co. Sec. Shelf
  1.2 Holding Co. Sec. Shelf

1.3 Aggregator Subsidiary: Correspondent, wholesale, retail
  1.4 Ind./Affil. Dep.
  1.5 Ind. MC
  1.6 Ind. Broker

2.0 IBank/Ind. Holding Company
  2.1 Holding Co. Sec Shelf
  2.2 Holding Co. Sec Shelf

2.3 Aggregator Subsidiary: Correspondent, wholesale
  2.4 Ind. Dep.
  2.5 Ind. MC
  2.6 Ind. Broker

Supply Chain
Network
Example: Trees – 2006 Mortgage Originations
Example: Networks – 2006 Mortgage Originations

► Develop an equilibrium model of intermediaries that share risk in a financial network.

► Key properties:

● Participation in network is voluntary (assume pairwise stability as in Jackson and Wolinsky, 1996)
● Agents decide whether to invest in quality
● *Financial norms* represented by investment decisions
● Model simple enough to allow for numerical approximation of equilibrium in mid-large-sized networks.

► Three general implications:

1. Network structure influences financial norms – both direct and indirect counterparties.
2. Heterogeneous financial norms often coexist in the network.
3. Network proximity is related to financial norms.
Key Results

- Decisions to share risk and to invest in screening are both non-monotonic in capital requirements (analytic).
- Heterogeneous screening (quality) policies arise naturally within lending networks, leading to clusters of high and low quality lending (analytic).
- High quality lenders tend to have more connections and these tend to be of higher quality (simulation).
- Ex post, a sub-network of concentrated, low quality lending channels is observable in the 2006 U.S. mortgage market (empirical).
- Systemically, pivotal firms can be forecast (empirical and analytical).
Pivotal Firm Linkages
No Single Data Set Links Lenders, House Prices with Time-Matched House Characteristics and Mortgage IDs

<table>
<thead>
<tr>
<th></th>
<th>DataQuick</th>
<th>Modified DataQuick</th>
<th>Corelogic</th>
<th>Trulia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transaction Prices</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>House Characteristics</strong></td>
<td>Static</td>
<td>Updated Annually</td>
<td>Static</td>
<td>Static</td>
</tr>
<tr>
<td><strong>Lien Recording History</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Foreclosure</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Recorded Mortgage Originator</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Mortgage ID</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>House Price Indices</strong></td>
<td>Repeat Sales/Hedonic</td>
<td>Dynamic Hedonic</td>
<td>Repeat Sales</td>
<td>Repeat Sales</td>
</tr>
</tbody>
</table>
Shortcomings of repeat-sales indices

 ► **Small sample size:** Ignores houses that sell just once during sample period.

 ► **Changing Quality/Quantity:** Assumes house characteristics remain constant (and that relative prices of each characteristic remain constant).
  - Adding a new bedroom will make us think prices of *unaltered* houses have gone up.

 ► **Volatility:** Volatility of index will underestimate volatility of individual house prices.
  - Construction of index automatically induces smoothing.
  - Even ignoring this, index is a (somewhat) diversified portfolio, not an individual house.

 ► The first-differencing required in repeat-sales estimators mechanically induces serial correlation depending on the timing of sales.

 ► Sample-selection bias: Are houses that sell representative of houses that don’t?
Problems with repeat-sales indices

Most houses don’t sell very often

► A house is only included if it sells more than once, so we throw out a huge fraction of house sales.

► In San Francisco between 2003–2012,
  ● 24,342 houses sold at least once.
  ● 20,778 (85.4%) sold only once, so are discarded.
  ● Index considers only 3,564 houses (14.6%)

► In Los Angeles between 2003–2012,
  ● 236,406 houses sold at least once.
  ● 194,375 (82.2%) sold only once, so are discarded.
  ● Index considers only 42,031 houses (17.8%)
Problems with repeat-sales indices 2

Changes in property characteristics

► In San Francisco, 2502 Leavenworth sold in 2003 and 2008.

- In 2003:
  - Square footage = 1,752
  - Total rooms = 5
  - Bathrooms = 1
  - Price = $1,503,000

- In 2008:
  - Square footage = 2,913
  - Total rooms = 8
  - Bathrooms = 3
  - Price = $5,500,000

► Changes in size (or quality) are wrongly counted as “returns”
Problems with Repeat-Sales indices 3

The index is not a house price!

► For mortgage valuation, stress testing, etc., we need the distribution of future house prices.

► In repeat-sales methodology, each period’s index growth is a constant.
  - No specification of inter-period dynamics.

► Even if we assume house prices follow (say) geometric Brownian motion, we can’t just use the volatility of the index.
  - Index levels are estimated (though standard errors are never shown).
  - Construction of index automatically induces smoothing.
  - Volatility of index will underestimate volatility of individual house prices.
  - Even ignoring this, index is a (somewhat) diversified portfolio, not an individual house.

► We also can’t use index properties to test for (e.g.) serial correlation in house prices.
  - Estimation procedure mechanically induces serial correlation in index.
    ✤ Even when there is none in individual house prices.

Objective to account for:
- Dynamics flexibly.
- Property characteristics.
- Macroeconomic variables.
- All sales, not just repeats.
- Unobserved heterogeneity across properties.

Purpose of the index
- Develop suitable dynamic house price index for applications in mortgage valuation.
- Develop suitable index for DFAST nine-quarter-ahead planning horizon – incorporate macro fundamentals directly into index.
New Index

Write log-price, $y_{i,t}$, as

\[
y_{i,t} = \underbrace{A_b x_{i,t}}_{\text{house price index}} + B_b \xi_t + \alpha_{b,t} + \mu_i + \epsilon_{i,t},
\]

\[
= X_t \beta_b + \alpha_{b,t} + \mu_i + \epsilon_{i,t},
\]

\[
\alpha_{b,t} = \rho \alpha_{b,t-1} + \eta_t,
\]

where

\[
\epsilon_{i,t} \sim \text{i.i.d. } N(0, \sigma^2_\epsilon),
\]

\[
\mu_i \sim \text{i.i.d. } N(0, \sigma^2_\mu),
\]

\[
\eta_{i,t} \sim \text{i.i.d. } N(0, \sigma^2_\eta).
\]

- $x_{i,t}$ = home hedonics (beds, size, lot size).
- $\xi_t$ = macro fundamentals (interest rates, population, unemployment).
- $\mu_i$ = house-specific random effect (due to unobservable differences).
- $\alpha_{b,t}$ = unexplained (and unobserved) portion of the index.
Model can be written as a standard, linear state-space model if we augment the state, $s_t$, to include $\mu_i$:

$$
\begin{pmatrix}
\alpha_{b,t} \\
\beta_{1,t} \\
\beta_{2,t} \\
\vdots \\
\beta_{k,t} \\
\mu_{1,t} \\
\mu_{2,t} \\
\vdots \\
\mu_{I,t}
\end{pmatrix} = \begin{pmatrix}
\alpha_{b,t} \\
\beta_{1} \\
\beta_{2} \\
\vdots \\
\beta_{k} \\
\mu_{1} \\
\mu_{2} \\
\vdots \\
\mu_{I}
\end{pmatrix} = \begin{pmatrix}
\alpha + \rho \alpha_{b,t-1} + \eta_t \\
\beta_{1,t-1} \\
\beta_{2,t-1} \\
\vdots \\
\beta_{k,t-1} \\
\mu_{1,t-1} \\
\mu_{2,t-1} \\
\vdots \\
\mu_{I,t-1}
\end{pmatrix}.
$$
In principle, we could estimate the model’s parameters by maximizing the likelihood function for this state space model using standard Kalman Filter (Kalman and Bucy, 1961).

We don’t do it this way.

- Dimensionality of problem is very high.
- Lots of missing data.
- Want to be able to extend to even more general settings.
  - E.g., more general distributional assumptions.

Instead we use Markov Chain Monte Carlo (MCMC) Bayesian methods (e.g., Johannes and Polson, 2009).

- Details in paper.
Alameda County: Comparison WRS Indices and Dynamic Index
San Diego County: Comparison WRS indices and Dynamic Index
San Francisco County: WRS indices and Dynamic Index
Alameda Zip = 94618: Comparison Zillow Medians, Medians and Model
San Diego Zip = 92129: Comparison Zillow Medians, Medians and Model
San Francisco, Zip = 94111: Comparison Zillow Medians, Medians, and Model
Conclusions

► Without a mortgage loan ID integrating existing data sets is a significant hurdle – regulatory oversight is challenging.

► Random sampling from non-integrated loan-level mortgage data sets may lead to biased estimates of the true population distributions of the contract space, the geographic coverage of the contracts, and industrial organization of the industry.

► Lack of dynamic house price data sets present significant challenges for disentangling price from quality changes in the housing market – especially a challenge in coastal states with high real option values.

  • Currently, market and regulators rely on WRS indices due to lack of data.
  • Use of a fundamentally static price indices leads to significant problems in mortgage valuation and stress testing.

► New large data set matching and learning algorithms hold considerable promise in addressing both the mortgage and the house price data challenges for economists and regulators.