

The Dearth and Life of Subprime Mortgage Data: An Overview of Data Sources for Market Modeling ¹

Vanessa G. Perry, PhD

The George Washington University
School of Business
2201 G Street, NW, 301 Fungler Hall
Washington, DC 20052
202-994-4190 Phone
202-994-8999 Fax
vperry@gwu.edu

Working Draft

January 8, 2008

¹ Important contributions to this overview were made by Jesse Abraham, Amy Crews-Cutts, Marsha Courchane, Jim Follain, Richard Green, Ellen Merry, Anthony Pennington-Cross, Maury Seldin, Brent Smith, and Bob Van Order.

The subprime mortgage market debacle has fueled a flurry of attention and debate among policymakers (especially Congress and regulatory agencies), researchers, lenders, investors, advocates and others. In recent months, a variety of proposed solutions have been put forth by financial regulators, Congress, and consumer organizations. These proposals range in scope, but are intended to help a select group of subprime borrowers avoid foreclosure. While it is clear that avoiding widespread foreclosures is an important outcome, it is unclear whether these proposed alternatives will have the intended impact—particularly on local housing markets. Because many aspects of subprime crisis are unprecedented, the effects of these subprime market developments are largely unpredictable. Neither the consequences nor the antecedents of the current subprime situation are well understood. While noting the lack of adequate data for empirical tests, Green and Wachter (2007) argue that the problems in the subprime market are due in large part to underlying information asymmetries between borrowers and lenders and originators and investors.² These information asymmetries have persisted because different market players—including lenders, servicers, investors and public policy analysts—rely on different data to inform their decisions. Without integrating information on borrower characteristics, loan characteristics, property information, loan performance, and house prices, it will remain impossible to understand or predict the effects that the current market situation will have on the housing market and the economy going forward.

In sum, there is a dearth of empirical evidence of the antecedents or consequences of the subprime crisis, largely due to a lack of data appropriate for model testing and

² Richard K. Green and Susan M. Wachter, “The Housing Finance Revolution,” presented at the Kansas City Federal Reserve’s 31st Annual Economic Policy Symposium, Jackson Hole Wyoming, August 31, 2007.

development. The purpose of this report is to provide an overview of extant datasets that have been used to analyze recent developments in the mortgage market, and to propose alternative data sources that would enable comprehensive econometric modeling of these phenomena.

The Dearth of Data

Concerns about the availability and quality of data on the subprime market are not new. Much of the early research on the subprime market has been based on analyses of Home Mortgage Disclosure Act (HMDA) data or proprietary lender data. There are well-documented drawbacks to both these data sources. Yezer (2004), for example, highlighted several limitations of HMDA coverage of the subprime market, and recommended the use of property records that contain information on mortgagees and mortgagors as well as transactions prices and property location that could be matched with other records, particular credit history, using the name of the mortgagor.³ Smith (2004) argued for better measures of the impact that subprime loans have on neighborhood foreclosure levels.⁴ Merry and Wilson (2007) and Follain and Sklarz (2008) point out that a significant amount of research that has been conducted to examine the factors that affect the performance of individual mortgage loans is based on proprietary data sources.

This report is organized as follows. The next section describes the research agenda and modeling approaches. This is followed by an overview of available data sources of

³ Yezer, Anthony M. (2004), Written Testimony, U.S. House of Representatives Committee on Financial Services, Subcommittee on Housing and Community Opportunity Subcommittee on Financial Institutions and Consumer Credit, March 30, 2004, http://www.chase.com/cm/chf/miscellaneous/file/document/yezer_hmda.pdf.

⁴ Testimony of Geoff Smith Project Director Woodstock Institute, Before the Subcommittee on Housing and Community Opportunity And Subcommittee on Financial Institutions and Consumer Credit, U.S. House of

publicly-available and proprietary loan-level, property and local area, and loan performance data. Another section is devoted to data sources that provide supplementary economic indicators by geographic area. Finally, implications of utilizing these data are discussed.

The ‘End Game’ Mortgage Market Model

The purpose of the proposed research agenda is to examine the impact of subprime lending on local and national housing markets. In particular, in order to inform industry practitioners as well as the policy agenda, it is important to understand the relationship between the subprime market, house prices, default and foreclosures, and other economic factors.

An ideal model would forecast house price downturns and identify areas at risk of increased foreclosures due to subprime activity. The effects of foreclosures on inventories and other market outcomes are perhaps the most urgent focus of this research. Identifying submarkets that are most vulnerable to a cascading decline of house prices would allow lenders and the government to engage in appropriate interventions.

This model would also analyze subprime loan quality and performance to predict the probability of default and foreclosure, as well as loss severity, based on loan type, terms and local economic and housing market conditions.

While it is true that housing markets are local, local housing market occurrences do affect other markets and the national housing market.⁵ While a national-level model is clearly

Representatives Regarding Subprime Lending: Defining the Market and its Customers Tuesday, March 30, 2004.

⁵ Smith, Brent (2007), “The Subprime Mortgage Market, a Review and Compilation of Research and Commentary” Homer Hoyt Institute, October 17, 2007, <http://www.hoyt.org/subprime/index.html>.

important, local area models may be even more important in addressing the overall purpose of this effort.⁶

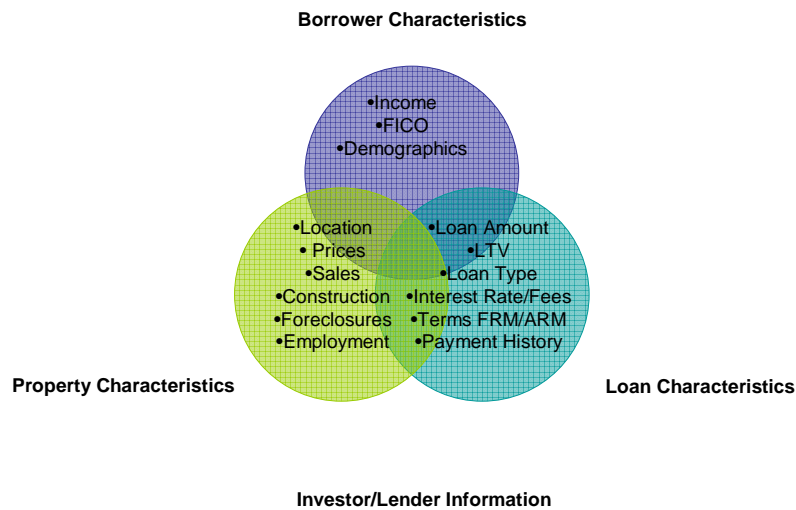
These modeling approaches build largely on previous research. There is an extensive published literature on modeling house prices and loan performance. However, these models predated the current structure on the mortgage market. Thus, these empirical approaches are ill-suited for the more recent, widespread incidences of subprime lending and securitizations that have occurred in an environment of unusual interest rate and house price patterns. Again, modeling of current market conditions has been limited owing to a lack of appropriate, available data.

Data needs for development of the models described above are summarized in Figure 1. According to this diagram, the ideal dataset would incorporate three main types of variables: Borrower characteristics; property characteristics, including local economic factors; and loan characteristics, including loan performance. In addition, data on lender, servicer and investor characteristics might serve as background variables, particularly as they relate to loan performance.

⁶ A new paper by Gabriel, Green and Redfearn (2008) implies that no one model can capture house prices across markets, or even across time within a market.

Figure 1

What Data Do We Need for Price and Performance Modeling?



Clearly there is no single source of data that includes all of these categories of variables. Therefore, compilation of this kind of dataset requires merging loan-level data with information available by geographic area, such as county, MSA or state. The next sections provide examples of publicly- and privately-available data sources that could be used as inputs to a comprehensive dataset.

LOAN DATA SOURCES

Home Mortgage Disclosure Act Data (HMDA)

HMDA data have been widely used to analyze mortgage lending patterns. The Home Mortgage Disclosure Act was enacted by Congress in 1975 and is implemented by the Federal Reserve Board's Regulation C. This regulation provides the public loan data that can be used to assist in determining whether financial institutions are serving the housing needs of their communities; to attract private investment to underserved areas; and to identify possibly discriminatory lending patterns.

This regulation applies to certain financial institutions, including banks, savings associations, credit unions, and other mortgage lending institutions. In 2007, there were approximately 34.1 million loan records for 2006 reported by 8,886 institutions. In 2006, 8,848 financial institutions reported approximately 36.4 million loan records for CY 2005 (FFIEC 2007).

In response to the growth of subprime lending, the Federal Reserve Board revised Regulation C in 2002 to require that lenders identify and disclose information about mortgages with annual percentage rates (APRs, which encompass interest rates and fees) above designated thresholds. Other new disclosures included lien status (whether a loan is a first lien, a junior lien, or unsecured home improvement loan) and whether a loan is secured by a manufactured home or is subject to the protections of the Home Ownership and Equity Protection Act of 1994 (HOEPA). This additional data was intended to facilitate the identification of subprime loans. Prior to the collection of this additional data, subprime loans could only be identified in Home Mortgage Disclosure Act (HMDA) data by matching

the reporting lending institution with a list of known subprime lenders updated annually and published by the Federal Reserve.⁷ A list of variables available on the HMDA dataset is provided in Appendix I.

These loan-level data are often matched with Census data by MSA or Non-Metro County. The HMDA-provided census dataset includes local area demographic, income, population, and housing variables from the most recent decennial Census (i.e., 2000). Analyses using these data have been invaluable for research on the growth and distribution of subprime lending in demographic and geographic terms (cites). Many analysts, including Yezer (2004) and Avery, Brevoort and Canner (2007), identify several important limitations of using HMDA data for analyzing the subprime market. First, certain small lenders, particularly non-depository institutions, are not required to file HMDA data. Another problem lies in the definition of subprime used in these data—currently measured by a flag for loans with APRs that are 3 percentage points above a comparable Treasury APR. According to Avery et al. (2007), some subprime loans do not meet this standard.⁸ Another criticism is that the 3 percentage point cutoff does not take into account the yield curve. Finally, the APR is not the same as the yield or effective cost of a mortgage, because it does not reflect fees or prepayment penalties—both common to subprime loans (Green and Wachter 2007).

HMDA studies of mortgage outcomes, particularly subprime analyses, are often criticized for omitted variable problems. Omitted variables include some of the most important factors lenders rely upon for underwriting and pricing, such as loan-to-value (LTV) ratios and measures of borrower credit history. Detailed loan terms, including fixed-

⁷ <http://www.huduser.org/datasets/manu.html>

⁸ Avery, Brevoort and Canner (2007), “The 2006 HMDA Data,” Federal Reserve Bulletin.

versus adjustable-rate or features such as negative amortization or prepayment penalties, are also not provided on the HMDA dataset.

Enhancing HMDA Data

Several researchers have circumvented these limitations by appending data from other sources to HMDA. For example, Avery et al. (2007) employ a proprietary database, TrenData⁹, which tracks loan performance from the credit records of a geographically-stratified random sample of about 30 million individuals over time. In particular, with these data these authors computed the proportion of mortgage borrowers in a county that are at least ninety days delinquent on their mortgages. These authors then examine the relationship between the rates of serious delinquency on mortgages and the incidence of higher-priced lending for 2005 and 2006 as well as county-level economic indicators including the unemployment rate, per capita income, house price appreciation, and population growth. Results suggest that house price appreciation was the most important economic factor in predicting future loan performance, but also that the incidence of higher-priced lending has independent predictive value for loan performance beyond that of economic factors.

Several recent studies have explored trends and patterns of foreclosures in particular geographic areas. These studies use databases of foreclosed properties, and tend to follow increases in the number of foreclosures in a particular location (Merry and Wilson 2007). For example, Immergluck and Smith (2004; 2005) used foreclosure data provided by a firm called Foreclosure Report of Chicago to matched with HMDA data to analyze the impact of

⁹ TrenData are a registered trademark of TransUnion LLC (<http://products.trendatatu.com/faqs.asp>)

subprime loans on foreclosure rates in various census tracts in the Chicago area.¹⁰ Firms that provide foreclosure data are plentiful, although they tend to specialize in particular metropolitan areas or states (for a list see http://www.digitaldeal.com/lists/foreclosure_lists.asp).

These examples demonstrate that HMDA data, while imperfect, can be useful for analyzing lending patterns if not for developing local as well as national market models.

The American Housing Survey (AHS)¹¹

The Department of Housing and Urban Development (HUD) and the U.S. Bureau of the Census conduct the biennial American Housing Survey (AHS). This survey collects information on physical characteristics of housing units, demographic characteristics of occupants, and neighborhood characteristics. In addition, financial characteristics including mortgage amount, home value, interest rate, mortgage type, loan purpose, taxes and other housing costs are reported. Unlike other housing and mortgage market surveys, the AHS is longitudinal survey and collects data from the same housing units every two years. The Census uses a multi-stage probability sampling methodology that results in a sample of 60,000 housing units and 55,000 households. Additional AHS surveys provide periodic examinations of the housing stock in 21 major metropolitan areas.

The AHS has not been used widely for research on the subprime market. The AHS questions on mortgage products are not user-friendly and do not identify many current forms of mortgages. Improvements to the survey instrument are under consideration.¹²

¹⁰ Immergluck, Dan and Geoff Smith(2004), “Risky Business – An Econometric Analysis of the Relationship Between Subprime Lending and Neighborhood Foreclosures,” Woodstock Institute; Dan Immergluck, Geoff Smith (2005), Measuring the effect of subprime lending on neighborhood foreclosures: Evidence from Chicago. *Urban Affairs Review*. Thousand Oaks: Jan 2005. Vol. 40, Iss. 3; p. 362

¹¹Details can be found at <http://www.huduser.org/datasets/ahs.html>.

Although the AHS has been criticized for under-representing important market segments and other errors, some research suggests that the AHS could be a very powerful resource for answering many housing finance research and policy questions. These micro-data have the potential to inform policy makers and the mortgage industry on important lending and market outcomes, due to its detailed household, housing, loan, and geographic information.¹³

First American LoanPerformance (LP) Data

LP securities data are perhaps the ‘Rolls-Royce’ of loan-level data. These proprietary data are widely-utilized by other institutions, and used by LP itself for default and loss modeling. LP maintains a database of over 100 million active and paid-off loans tracked for delinquency, prepayment, and foreclosure. The company also owns the nation’s largest servicing performance data repository, and a database of loan-level details on non-agency, publicly-traded mortgage and asset-backed securities. Most importantly, LP has compiled the largest credit and prepayment performance database of subprime loans. Appendix II includes a list of key variables provided by LP.¹⁴

LP data are used broadly. For example, Moody’s uses detailed performance history data from LP to model the performance of loans and securitizations, and to measure the relationship between economic characteristics and loan behavior.¹⁵ With these data,

¹² “The American Housing Survey and Non-Traditional Mortgage Products,” U.S. Department of Housing and Urban Development Office of Policy Development and Research, Frederick J. Eggers and Donald Bradley, Econometrica, Inc., September 2007.

¹³ “Analysis of Housing Finance Issues Using the American Housing Survey (AHS),” U.S. Department of Housing and Urban Development, PD&R, Ken Lam and Bulbul Kaul, Abt Associates Inc., April 2003.

¹⁴ “Company Overview,” <http://www.loanperformance.com/infocenter/literature.aspx#documentation>.

¹⁵ Siegel, Jay (2003), “Moody’s Mortgage Metrics: A Model Analysis of Residential Mortgage Pools,” <http://www.moody.com/cust/content/Content.ashx?source=StaticContent/Free%20Pages/Products%20and%20Services/Downloadable%20Files/m3%20special%20report.pdf>.

Moody's forecasts the future performance of prime and subprime loans under a large number of different projected future economic scenarios. The foundation of the forecast includes credit bureau scores, the amount of equity borrowers have in their homes, how fully the borrowers documented their income and assets, whether the borrower intends to occupy or rent the property¹⁶, and whether the loan is for purchase or for refinance.¹⁷

LP data have also been used for scholarly research. Danis and Pennington-Cross (2005) utilized LP data matched with OFHEO House Price Index (HPI) data, Bureau of Labor Statistics data on the unemployment rate by state and data on interest rates from the Freddie Mac's Primary Mortgage Market Survey to model subprime delinquency, default, and prepayment. In addition to significant effects of borrower FICO score and LTV, these authors found evidence that the change in interest rates affects prepayment, default, and delinquency. They also found that prepayment penalties extend the duration of subprime loans, and documentation status is associated with delinquency.¹⁸

In a recent paper, Merry and Wilson (2007) use LP data combined with state-level house price data from OFHEO and state-level credit quality data from TransUnion's Trendata to examine subprime and prime mortgage default rates across the country, and geographic variations therein. Using this combined dataset, these authors test for effects of local economic conditions, borrower characteristics, loan terms, and lender behavior on default.¹⁹

¹⁶ This variable is not well measured. Not all borrowers are truthful about intended use, and some borrowers begin by living in a home, and then move elsewhere and rent out the home.

¹⁷ Testimony of Warren Kornfeld, Managing Director Moody's Investors Service Before the Subcommittee on Securities, Insurance and Investment United States Senate, April 17, 2007.

¹⁸ **The Delinquency of Subprime Mortgages** Michelle A. Danis and Anthony Pennington-Cross (2005) Working Paper 2005-022A <http://research.stlouisfed.org/wp/2005/2005-022.pdf>.

¹⁹ **The Geography of Mortgage Delinquency**, Ellen A. Merry and Michael D. Wilson, Federal Reserve Board of Governors, Presented at the AREUEA Annual Meetings, Chicago, IL, January 6, 2007.

Access to the best data for these purposes (LP) is restricted to lending institution subscribers, although access has been granted on a limited basis to regulators. LP data includes only loans that have been securitized, and suffers from truncation problems since loans and pools that prepay or otherwise terminate are not represented.²⁰

Applied Financial Technology (AFT)-Fidelity National Information Services

Applied Financial Technology²¹ maintains a loan-level database of non-Agency-Collateralized Mortgage Obligations (CMO) collected from trustees. These data represent approximately 70 percent of CMOs from the past 5 years. The firm has been recently purchased by Fidelity National Information Services, Inc., a leading provider of data processing for financial institutions, including mortgage loan processing and mortgage-related information products. Fidelity also collects county records data for approximately 80 percent of properties. The combined database contains detailed loan characteristics, property information and monthly performance data. A detailed list of available records is available in Appendix III.

AFT data are similar to the LoanPerformance securities data, although its coverage is less complete. In addition, AFT data are reportedly less expensive.

McDash Analytics

McDash collects loan-level information from seven of the top 10 residential mortgage servicers in the industry. The McDash database contains loan characteristics and

²⁰ Subprime Refinancing: Equity Extraction and Mortgage Termination Anthony Pennington-Cross; Souphala Chomsisengphet, *Real Estate Economics*; Summer 2007; 35, 2, pg. 233-63; Chomsisengphet, S. and A. Pennington-Cross (2006), "The Evolution of the Subprime Mortgage Market," Federal Reserve Bank of St. Louis Review, 88 (1), 31-56.

performance variables, including 12 years of historical data.²² In particular, loan and underwriting details, property zip location, property characteristics, and performance data are provided. A more detailed list of variables available on this dataset is provided in Appendix IV.

McDash data are similar to LP with a few important exceptions. LP data are collected from trustees, and includes more coverage of subprime loans—reportedly more than 80 percent of the market. McDash data, on the other hand, are collected from servicers and includes approximately 40 to 50 percent of the subprime market (about 5 million subprime loans originated since 1989). However, McDash is currently increasing its subprime data coverage through the HopeNow²³ alliance, a group of mortgage servicers formed to establish best practices for the loss mitigation in the subprime market. Also, there is no way to identify securities in McDash.

Moody's

Moody's Investors Service is a well-known source for credit ratings, research and risk analysis. In rating a subprime mortgage-backed securitization, Moody's first estimates the amount of cumulative losses that the underlying pool of subprime mortgage loans are expected to suffer over the lifetime of the loans. Moody's bases its estimates on loan-level data that it collects; among the variable its collects are credit bureau scores, borrower equity, income and asset documentation, owner occupancy, and loan purpose. Moody's also collects data on underwriting standards and past performance of similar loans made by that

²¹ <http://www.aftgo.com/>

²² <http://www.mcdash.com/products/>

²³ "Subprime: 'HopeNow' may help only so much," Jeanne Sahadi, CNNMoney.com, http://money.cnn.com/2007/10/12/real_estate/mortgage_alliance/index.htm?eref=yahoo

lender, the representations and warranties the lender is willing to provide regarding the loans, and servicer history. Moody's does not, however, receive any loan-level information that identifies the borrower or the property.²⁴

Lender Data

Lending institutions collect detailed information on mortgage originations and purchases, including loan and property characteristics, and these sources can be used to develop loan performance models. In a recent example, Courchane (2007) used a loan-level dataset provided by an anonymous group of lenders to analyze demographic differences in loan pricing. The proprietary lender data included FICO score, LTV, and debt-to-income ratios, contract rate type, loan purpose, occupancy status, and documentation type, and loan amount. These data were merged with census-tract level HMDA data to measure market competition, and also merged with Census data for tract level demographic characteristics.²⁵ In another example, Alexander et al. (2002) used a sample of over 23,000 loans originated or acquired by a national subprime mortgage lender to examine the relationship between third-party originations and default.²⁶

²⁴ Testimony of Warren Kornfeld, Managing Director Moody's Investors Service, Before the Subcommittee on Securities, Insurance and Investment United States Senate, April 17, 2007.

²⁵ The Pricing of Home Mortgage Loans to Minority Borrowers: How Much of the APR Differential Can We Explain? Marsha J Courchane. *The Journal of Real Estate Research*. Oct-Dec 2007. Vol. 29, Iss. 4; pg. 399-440.

²⁶ Some Loans Are More Equal than Others: Third-Party Originations and Defaults in the Subprime Mortgage Industry, William P Alexander, Scott D Grimshaw, Grant R McQueen, Barrett A Slade. *Real Estate Economics*. Winter 2002. Vol. 30, Iss. 4; pg. 667, 32 pgs.

PROPERTY DATA SOURCES

Public records data can serve as alternative to mortgage data collected by lending institutions.²⁷ For example, these data include important factors such as loan to value ratio, interest rate at origination, local area lending activity, loan type, and prepayment purpose. Follain and Sklarz (2008) used public records data from Fidelity Information Services²⁸ to model mortgage terminations, including foreclosure and refinancing, and after matching local house price data were able to test these models at the zip code-level.²⁹

Similarly, DataQuick provides ShareData, which is a commercial database of nationwide mortgage loan and sales information. ShareData provides retail and wholesale lending information, equity loan/credit line information, and local area demographic profiles. ShareData information is collected from county recorders and county tax assessors from across the nation on a loan- or property-level basis. For example, DataQuick sells loan-level data including property address, loan amount, purpose, loan type, property value, interest rate, and lender for prime and subprime loans. These data can be matched with other data, such as house prices, foreclosure activity, or other local economic variables based on property location. Lax, et al. (2004)³⁰ and Walters and Hermanson (2001)³¹ used the DataQuick database to draw a sample of borrowers for a survey on subprime lending practices.

²⁷ U.S. Mortgage Borrowing: Providing Americans with Opportunity, or Imposing Excessive Risk? An Empirical Analysis of Recent Foreclosure Experience in U. S. Mortgage Lending and Particularly Subprime Lending, May, 2007, The Center for Statistical Research Inc., Alexandria, Virginia.

²⁸ http://www.fidelityinfoservices.com/NR/RDONLYRES/51C91159-E8BD-4050-A835-FF84F70D6B05/0/DS04_DATASERVOVER.PDF

²⁹ Evaluating Mortgage Performance with Public Records Data, Presentation at 2008 AREUEA Session, January 4, 2008, Jim Follain and Mike Sklarz

³⁰ Lax, Howard, Michael Manti, Paul Raca and Peter Zorn (2004), "Subprime Lending: An Investigation of Economic Efficiency," *Housing Policy Debate*, 15(3), 533-71.

³¹ Subprime Mortgage Lending and Older Borrowers, *Research Report*, Neal Walters and Sharon Hermanson, AARP Policy Research Institute, March 2001.

Supplementary Data Sources

Loan level and public record data can be matched by geography with supplementary economic indicators. These include the Census (www.census.gov), loan performance and credit profiles by metropolitan area (creditforecast.com), house price appreciation (e.g. OFHEO, CMHPI), delinquency and foreclosure activity (MBA), and interest rates (Freddie Mac PMMS). These data, widely-used in previous studies, are described in Appendix IA and Appendix IV.

WHAT NOW?

In order to develop the models necessary to predict the outcomes of the subprime mortgage situation, researchers need access to quality data on subprime borrowers, property and area factors, and loan characteristics. Although it has the drawback of being costly and proprietary, LP is perhaps the most comprehensive data set for studying the subprime market. Alternative sources include McDash, proprietary lender data, property data drawn from public records matched with commercially-available foreclosure data, and the publicly-available HMDA dataset.

One approach to filling the loan-level data gap is to recruit a consortium of lenders that would be willing to contribute data to pool for the purpose of better understanding the subprime crisis..

The limitations of HMDA analysis have been well-documented, yet these data are useful owing to their widespread coverage of the market, not to mention their low cost, longstanding research tradition, and widespread availability.

There appears to be substantial, untapped potential of public records data combined with other data sources for developing market models. These kinds of data have only recently been maintained on a widespread basis in electronic formats. The limitations, of course, stem from a lack of detail on borrower characteristics (e.g., income and FICO), although these can potentially be overcome with LTV and CLTV estimates and local area demographic information.

Each of these approaches require extensive resource commitments— both financial and in terms of human capital. Yet the only way to pursue the subprime research agenda at hand is to identify data alternatives. This vexing problem may be the reason we lack sufficient proposed rigorous research on this very important issue. One way to encourage and expedite high-quality subprime research might be to dedicate resources upfront to the compilation of one or two datasets. Providing these data will undoubtedly spark the kind of urgent execution necessary to understand the effects of subprime on the local and national economy.

Appendix I Home Mortgage Disclosure Act Data (HMDA)³²

The Federal Reserve Board's Regulation C requires lenders to report the following information on home-purchase and home-improvement loans or refinancings.

For each application or loan:

- application date and the date an action was taken on the application
- action taken on the application
 - approved and originated
 - approved but not accepted by the applicant
 - denied (with the reasons for denial—voluntary for some lenders)
 - withdrawn by the applicant
 - file closed for incompleteness
- pre-approval program used (for home-purchase loans only)
- loan amount
- loan type
 - conventional
 - insured by the Federal Housing Administration
 - guaranteed by the Veterans Administration
 - backed by the Farm Service Agency or Rural Housing Service
- pre-approval status
- lien status
 - first lien
 - junior lien
 - unsecured
- loan purpose
 - home purchase
 - refinance
 - home improvement
- type of purchaser (if the lender subsequently sold the loan)

For each applicant or co-applicant:

- race
 - ethnicity
 - sex
 - income relied on in credit decision
- 2006 HMDA Data -56- 2006 HMDA Data
- For each property*
- location, by state, county, and census tract
 - type of structure
 - one-to four-family dwelling
 - manufactured home
 - multifamily property (dwelling with five or more units)
 - occupancy status (owner occupied or non-owner occupied)

For loans subject to price reporting:

- spread above comparable treasury security

For loans subject to HOEPA:

- indicator of whether loan is subject to HOEPA

In addition, information is also reported on home loans purchased by an institution during the calendar year.

³² <http://www.ffiec.gov/hmdafeedback/hmdaproducts.aspx>

Appendix IA Selected Variables from the Census³³

Summary Census Demographic Information

Tract Income Level - This corresponds to tract classifications as defined by the HMDA and CRA regulations. This field is based on the Tract Median Family Income %:

If the Median Family Income % is < 50% then the Income Level is Low.

If the Median Family Income % is >= 50% and < 80% then the Income Level is Moderate.

If the Median Family Income % is >= 80% and < 120% then the Income Level is Middle.

If the Median Family Income % is > =120% then the Income Level is Upper.

If the Median Family Income % is 0% then the Income Level is Not Known.

Tract Median Family Income % - This compares the tract level Median Family Income (MFI) to the MA level MFI. It is calculated by dividing the 2000 Tract MFI by the 2000 MA MFI. (For tracts located outside of an MA, the MFI used in the denominator is the statewide non-MA MFI. This figure is calculated using incomes from all areas of a state that are not assigned to MAs). Go to Tract Income Level to see how the tracts are grouped based on their Median Family Income %.

2000 HUD MA Median Family Income - This is the estimated Median Family Income as determined by the U.S. Department of Housing and Urban Development (HUD) for each year. HUD annually estimates this income for each MA in the United States and for each state's non-MA area.

2000 Est. Tract Median Family Income - This is the census tract's estimated Median Family Income (MFI) for each year, based on the HUD estimate for the MA or non-MA area where the tract is located. It is calculated by multiplying the Tract's Median Family Income % by the annual HUD MA MFI.

2000 Tract Median Family Income - This is the census tract's Median Family Income based on the 2000 Census. This field is called Tract Median Fam Inc on the CD-ROM.

Tract Population - This is the census tract's population based on the 2000 Census.

Tract Minority % - This is the percentage of the tract's total population minus white alone population according to the 2000 Census. It is calculated by dividing the Minority Population by the Tract Population.

Minority Population - This is the census tract's total population minus white alone population based on the 2000 Census.

³³ <http://www.ffiec.gov/hmda/censusproducts.htm>

Owner-Occupied Units - This is the number of housing units that were lived in by the owner of the unit. This number was determined by the 2000 Census. This field is called Tenure Owner Occ Housing Units on the CD-ROM.

1- to 4-Family Units - This is the number of housing units that hold less than 5 families. Condominiums, townhouses, and single-family homes are included in this number. Apartment buildings are not included in this number. The number was determined by the 2000 Census. This field is called Units Structure 1 to 4 subtotal on the CD-ROM.

Census Population Information

Tract Population - This is the census tract's population based on the 2000 Census.

Tract Minority % - This is the percentage of the tract's total population minus white alone population according to the 2000 Census. It is calculated by dividing the Minority Population by the Tract Population.

Number of Families - This is the number of families in the tract as determined by the 2000 Census. A family is defined as two or more related people living together.

Number of Households - This is the number of households in the tract as determined by the 2000 Census. A household is any residence, including those occupied by single people and unrelated groups of two or more. By definition, all families are also considered households, but not all households are families. This field is called HH Inc Total on the CD-ROM.

White Population - This is the number of people in the tract that listed themselves as non-Hispanic white on the 2000 Census.

Tract Minority Population - This is the census tract's total population minus white alone population based on the 2000 Census.

American Indian Population - This is the number of people in the tract that listed themselves as non-Hispanic American Indian or Alaskan Native on the 2000 Census.

Asian/Hawaiian/Pacific Islander Population - This is the number of people in the tract that listed themselves as non-Hispanic Asian or Pacific Islander on the 2000 Census.

Black Population - This is the number of people in the tract that listed themselves as non-Hispanic black/African/American on the 2000 Census.

Hispanic Population - This is the number of people in the tract that listed themselves as Hispanic only on the 2000 Census.

Other Population/Two or More Races - This is the number of people in the tract that listed themselves as non-Hispanic some other race plus two or more races on the 2000 Census. This field is called Nonhisp Combo Other Race Pop on the CD-ROM.

Census Housing Information

Total Housing Units - This is the total number of occupiable housing units in the tract as determined by the 2000 Census.

1- to 4-Family Units - This is the number of housing units that hold less than 5 families. Condominiums, townhouses, and single-family homes are included in this number. Apartment buildings are not included in this number. The number was determined by the 2000 Census. This field is called Units Structure 1 to 4 subtotal on the CD-ROM.

Median Age of Housing Stock - This is the median age of all housing units in the tract as determined by the 2000 Census. Remember that this would be the median age in 2000 so the median age now would be higher, unless a lot of new residences have been built since then.

Inside Central City? - This is a Yes/No indicator of whether this tract is part of the MA's central city.

Owner-Occupied Units - This is the number of housing units that were lived in by the owner of the unit. This number was determined by the 2000 Census. This field is called Tenure Owner Occ Housing Units in the CD-ROM.

Renter-Occupied Units - This is the number of housing units that are occupied by someone other than the owner. This number was determined by the 2000 Census. This field is called Tenure Renter Occ Housing Units on the CD-ROM.

Vacant Units - This is the number of housing units that are not occupied. This number was determined by the 2000 Census. This field is called Vac Status Vacant Housing Units on the CD-ROM.

Owner-Occupied 1- to 4-Family Units - This is the number of housing units that have less than 5 residences and were occupied by the property owner. It was determined by the 2000 Census. This field is called Tenure Own Occ 1 to 4 subtotal on the CD-ROM.

Appendix II

Selected Variables from LoanPerformance³⁴

Loan Characteristics

- Current and Original LTV, CLTV
- Current and Original Balance
- Current and Original Coupon
- Current Delinquency Status
- Property and Loan Type
- Purpose
- Occupancy Status
- Documentation
- First Payment Date
- Term
- Prepayment Penalty
- MI Coverage
- Geographic Location

ARM and 2nd Liens

- Margin, Reset Dates
- Floors, Ceilings, Caps
- Index, Neg. Am
- Balance on 1stlien

Borrower Information

- Credit Score
- Loan Grade
- Payment History

Market Information

- Interest Rates
- House Prices

³⁴ <http://www.loanperformance.com/>

Appendix III

Applied Financial Technology Selected Variables

- ObjectId
- ClassId
- Name
- Description
- Created
- LastModified
- OwnerId
- LastModifiedBy
- IndexValue
- MaxNumLoans
- Issuer
- LoanProgram
- Agency
- OrigDate
- OrigTerm
- CollateralCode
- AmortTerm
- OrigBal
- OrigLTV
- CurWac
- CurNet
- CurWALA
- CurWAM
- CurFactor
- CurBal
- CurLTV
- CurLoanSize
- LoanSize
- PrimaryResRatio
- SingleFamRatio
- FullDocRatio
- CashOutRatio
- State
- ZipCode
- BorrowerFICO
- CurBorrowerFICO
- BorrowerIncome
- CoBorrowerIncome
- CoBorrowerFICO
- PropertyTypeCode
- OccupancyCode
- GeoAreaCode
- LoanStatusCode
- ProductCode
- DocumentationTypeCode
- LoanPurposeCode
- RemittanceCode
- LoanCategoryCode
- SourceCode
- BankruptcyCode
- ForeclosureCode
- PayScheduleCode
- InvestorCode
- AFTRating
- SPRating
- MoodyRating
- FitchRating
- EspScoreRefi
- EspScoreHT
- EspScoreElbow
- Servicer
- ServicerNumber
- IndexType
- GrossMargin
- CouponResetPeriod
- LifeCap
- ResetCap
- LifeFloor
- ServicingFee
- TeaserPeriod
- TeaserRate
- PayResetPeriod
- PayResetCap
- PayCapWaivePeriod
- InitialPayResetPeriod
- NegativeAmortFlag
- NegativeAmortLimit
- IndexLookback
- LoanIDNumber
- LoanIDString
- PoolNumber
- AsOfDate
- DateModified
- UserModified

- CurNumLoans
- CPR1Month
- CPR3Month
- CPR6Month
- CPR12Month
- CPRLife
- IsSingleLoan
- IsARM
- IsABS
- HasPenalty
- IsMegaPool
- PreAggTypeID
- IsSecuritized
- RowId
- DealID
- PoolID
- OrigLTVRatio
- CombinedLTV
- PropertyValue
- CreditGrade
- FirstPayDate
- MaturityDate
- PrepayPenaltyTerm
- PrepayPenaltyCalc
- IsBalloon
- IsInterestOnly
- IOTerm
- IORules
- IsConvertible
- OrigPI
- FirstRateResetDate
- NextRateResetDate
- NextPaymentChangeDate
- GroupNumber
- MSACode
- LoanLevelLoanProgram
- OrigHPI
- EspScoreDel
- EspScoreFc

Appendix IV
McDash Analytics Core Database Selected Variables³⁵

- Origination Date
- Loan Closing Date
- First Payment Date
- Current Maturity Date
- Original Term
- Original Loan Amount
- Property Type
- Property State
- Property ZIP
- Property Value
- Occupancy
- FICO/Credit Score
- Underwriting Type
- DTI (Front End)*
- Mortgage Type
- Product Type
- Special Product
- Guarantee Fee*
- Net Service Fee
- Pass-Through Rate*
- Interest Type
- Loan Purpose
- P & I Frequency
- Loan Source
- Investor
- Remittance Type*
- Document Type
- MI Company
- Recourse Flag
- IO Flag Alpha
- Balloon Flag
- Balloon Term
- Pre-Payment Penalty Flag
- Pre-Payment Penalty Term
- Current Interest Rate
- Current FICO/Credit Score
- Principal Balance (This Loan)
- Next Payment Due Date
- Remaining Term (Amortization)
- Monthly Escrow Payment Amount
- Monthly Principal & Interest
- Loss Mitigation Flag
- Foreclosure Flag
- Bankruptcy Flag
- Bankruptcy Chapter Type
- Paid In Full Integer
- ARM Interest Reset Period
- ARM Index Alphanum
- ARM Initial Rate
- ARM Margin
- ARM Initial Rate Adjustment Period
- Option (ARM) Flag

*Restricted access

³⁵ <http://www.mcdash.com/products/>

Appendix V Supplementary Data

- **Moody's CreditForecast.com**

CreditForecast.com is a joint product of Equifax and Moody's Economy.com.

CreditForecast provides history for a wide range of household credit, economic, and demographic variables at a detailed level of geography. Users are able to examine, segment, and stratify credit risk and economic data across states, the top 200 metropolitan areas, and rest of states using quarterly historical credit data back to 1998.³⁶ The historical credit data are drawn from a 5% independent random sample of the Equifax National Consumer Database, which contains information collected from over 60,000 contributors who furnish data on tradelines, collections, public records and demographics across a broad range of industries.

A list of available MSA-level credit, housing and economic variables from CreditForecast is provided on their website.³⁷

- **OFHEO House Price Index**

The HPI is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties. This information is obtained by reviewing repeat mortgage transactions on single-family properties whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac since January 1975.

³⁶ www.creditforecast.com

³⁷ Source: http://www.creditforecast.com/credit_series.asp?active_div=tab_mort

The HPI serves as a timely, accurate indicator of house price trends at various geographic levels. Because of the breadth of the sample, it provides more information than is available in other house price indexes. It also provides housing economists with an improved analytical tool that is useful for estimating changes in the rates of mortgage defaults, prepayments and housing affordability in specific geographic areas.

The HPI includes house price figures for the nine Census Bureau divisions, for the 50 states and the District of Columbia, and for Metropolitan Statistical Areas (MSAs) and Divisions.³⁸

- **Freddie Mac CMHPI**

Freddie Mac's Conventional Mortgage Home Price Index (CMHPI) provides a measure of typical price inflation for houses within the US. Values are calculated quarterly at five levels of geographical aggregation: Metropolitan Statistical Area (MSA), state, Census division, and national. For the 11 largest metropolitan areas, the MSA and the Metropolitan Statistical Area Divisions (MSADs) that make up the MSA indices are available. Starting with the first quarter 2007 release of the CMHPI, Freddie Mac made available a purchase-transactions only series of the CMHPI in addition to the classic CMHPI series for the 9 Census divisions and the nation.³⁹

- **MBA Delinquency Survey**

The National Delinquency Survey is one of the most recognized sources for residential mortgage delinquency and foreclosure rates. Based on a sample of more than 44 million mortgage loans serviced by mortgage companies, commercial banks, thrifts, credit unions

³⁸ http://www.ofheo.gov/hpi_download.aspx

and others, NDS provides quarterly delinquency and foreclosure statistics at the national, regional and state levels.

Delinquency and foreclosure measures are broken out into loan type (prime, subprime, VA and FHA) and fixed and adjustable rate products. At each geographic classification, there are 7 measures: total delinquencies, delinquency by past due category (30-59 days, 60-89 days and 90 days and over), new foreclosures, foreclosure inventory, and seriously delinquent. The total number of loans serviced each quarter, as compiled through the survey, is also included in the data.⁴⁰

- **Freddie Mac PMMS**

Freddie Mac's Primary Mortgage Market Survey® (PMMS®) surveys lenders each week on the rates and points for their most popular 30-year fixed-rate, 15-year fixed-rate, 5/1 hybrid amortizing adjustable-rate, and 1-year amortizing adjustable rate mortgage products. The survey is based on first-lien prime conventional conforming mortgages with a loan-to-value of 80 percent. In addition, the adjustable-rate mortgage (ARM) products are indexed to constant-maturity U.S. Treasury rates and lenders are asked for the both the initial coupon rate and points as well as the margin on the ARM products.⁴¹

³⁹ <http://www.freddiemac.com/finance/cmhpi/>

⁴⁰ <http://www.mortgagebankers.org/ResearchandForecasts/ProductsandSurveys/NationalDelinquencySurvey.htm>

⁴¹ http://www.freddiemac.com/pmms/pmms_archives.html