CREDIT RISK, INVESTOR BEHAVIOR AND RESIDENTIAL MORTGAGE DEFAULT

LUO CHENXI

(B.A. Economics, Nankai University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF REAL ESTATE

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Luo Chenxi

LUO CHENXI

7 NOVEMBER 2014

Acknowledgements

First, I would like to express my deep gratitude to my thesis supervisor, Professor Deng Yongheng, for his kind guidance, great patience and persistent support throughout my entire PhD study. This thesis and all my potential achievements in academic and in life during the past few years would be impossible without his great support at various occasions. As a well-established econometrician and real estate economist, his great enthusiasm to academia, ultimate pursuit of academic breakthrough, and extreme hardworking deeply inspire me through these years.

I am also grateful to Associate Professor An Xudong at San Diego State University. I have been working with him for several research projects since my first year in my PhD life. He is not only a great scholar, but also a good friend and mentor. His kind support, patient guidance and optimism have greatly enhanced my research capacities and helped me through difficulties. Besides, I would like to thank my thesis committee members Professor Sumit Agarwal and Assistant Professor Qian Wenlan, for their extreme support and constructive suggestions on my research and my future career. The cooperating experience with them is always an invaluable treasury to me. Special thanks to Associate Professor Tu Yong for supporting me by offering a part-time research job when I got no scholarship in the first semester, and to Associate Professor Yu Shi Ming for hiring me as his research assistant and guiding me on the energy efficiency research with meticulous attitude.

I also want to express my thanks to Associate Professor Fu Yuming, Professor Ong Seow Eng, Associate Professor Sing Tian Foo, Associate Professor Zhu Jieming, Associate Professor Liao Wen-Chi and Dr. Lee Kwan Ok for their dedications to my coursework teaching and/or valuable comments on my thesis. Meanwhile, I benefit a lot from discussions with the research staff from the Institute of Real Estate Studies (IRES), e.g., Associate Professor Wu Jing from Tsinghua University, Dr. Li Pei, Dr. Zhao Daxuan, Dr. Ren Chaoqun, Dr. He Jia and Ms. Zhou Xuefeng. My postgraduate peers at the Department of Real Estate have also offered great help to my personal as well as research life in Singapore. This group includes but is not limited to: Li Zhuo, Zhang Bochao, Rengarajan Satyanarain, Li Zhiliang, Wang Yourong, Wang Yonglin, Xu Mingying, Lai Xiongchuan, Li Qing, Guo Yan, Qiu Leiju, Zhang Liang, and *etc*. I appreciate Zainab Bte Abdul Ghani, Ko Chen, Nor'Aini Binte Ali, Zheng Huiming and Li Congmiao for their administrative and logistic support through my PhD life.

I gratefully appreciate National University of Singapore for four years' the generous financial support. In addition, I want to thank the participants at the 2013 AsRES/AREUEA International Conference, 2013 China International Conference in Finance, and 2014 American Financial Association Conference, for their valuable comments on my papers.

Last but not least, I thank my family and friends in China and Singapore, who have always been there for me through ups and downs in life. In particular, I would never thank enough to my parents – Luo Binghua (罗炳华) and Chen

Youlan (陈友兰) for their unconditional love and persistent support. They have always been supportive in all my pursuits. Any achievement I've got so far, if any, is not possible without them. Therefore, this thesis is largely dedicated to them!

Table of Contents

Acknowledgements i

Table of Contents iv

Summary ix

List of Tables xi

List of Figures xiii

Chapter 1 Introduction 1

1.1 Background and Problem Statement 1

1.2 Objectives and Research Questions 13

1.3 Significance of the Research 18

1.4 Summary and Organization of the Thesis 23

Chapter 2 Literature Review 26

2.1 Introduction 26

2.2 General Review on Mortgage Defaults 28

2.2.1 Mortgage default process 28

2.2.2 Development of the default theory 29

2.2.3 Default risks in condominium market 35

2.3 Recent Financial Crisis: Triggers 36

2.3.1 Innovation in mortgage products 37

2.3.2 Fast growth of securitization 37

2.3.3 The wisdom of market players 38

2.3.4 Borrower (investor) behavior 38

2.4 Neighborhood Effects of Foreclosure and Foreclosure Concentration 42

2.4.1 General Literature on Neighborhood Foreclosure effects 43

2.4.2 Impact of Concentrated Foreclosures on borrowers' default decision 47

2.5 Probabilistic Data Linkage Methodology 49

2.5.1 General literature on Data linkage 49

2.5.2 Propensity score matching 53

2.5.3 Machine Learning 55

2.6 Summary 58

Chapter 3 The Hidden Peril: The Role of the Condo Loan Market in the Recent Financial Crisis 62

3.1 Introduction 62

3.2 Data Description 70

3.2.1 Data sources 70

3.2.2 Descriptive statistics of condo versus single-family loans 73

3.3 Hypotheses and Methodology 80

3.3.1 Hypothesis Development 80

3.3.2 Methodology 81

3.4 Empirical Analysis on Condo Default 82

3.4.1 Unconditional result of the default behavior of the condo market 82

3.4.2 Regression analysis of condo loan default behavior 84

3.5 What Drives the Default Pattern in the Condo Loan Market? 90

3.5.1 Unobserved characteristics of the condo market 90

3.5.2 Lender or borrower effect 94

3.5.3 Disentangling competing explanations: further evidence using Freddie Mac data 97

3.5.4 Why are condo borrowers riskier? 101

3.6 The Aggregate Implications of Condo Loan Defaults 107

3.6.1 Within-one-year default analysis 108

3.6.2 Do (early) condo defaults predict the single family subprime market's subsequent default rate? 110

3.7 Summary 116

Chapter 4 Foreclosure Concentration and the Exercise of Mortgage Default Options 119

4.1 Introduction 119

4.2 Data and Measures of Foreclosure Concentration 124

4.2.1 Data Sources 124

4.2.2 Measures of Foreclosure Concentration 127

4.2.3 Descriptive statistics 131

4.3 Hypotheses and Methodology 135

4.3.1 Hypothesis Development 135

4.3.2 Methodology 140

4.4 Empirical Analysis on the Impact of Foreclosure Concentration on Borrower Delinquency 142

4.4.1 The impact of foreclosure concentration on borrower default option 142

4.4.2 Impact of foreclosure concentration by different regimes ... 147

4.4.3 Impact of foreclosure concentration by different borrower groups ... 150

4.4.4 Impact of foreclosure concentration by different neighborhoods ... 154

4.5 Summary ... 162

Chapter 5 Probabilistic Data Linkage in Real Estate Studies: Applications of Propensity Score Matching and Hard Matching with Machine Learning Techniques ... 164

5.1 Introduction ... 164

5.2 Data Description ... 171

5.2.1 BlackBox Logic (BBX) ... 172

5.2.2 Home Mortgage Disclosure Act (HMDA) ... 173

5.2.3 Data Preparation and Harmonization ... 176

5.3 Methodology ... 179

5.3.1 Statistical hard matching (Method 1) ... 179

5.3.2 Machine learning techniques (Method 2) ... 180

5.3.3 Propensity score matching (Method 3) ... 188

5.4 Empirical Results on Data Linkage ... 191

5.4.1 Basic linking results of Method 1 and Method 2 ... 191

5.4.2 Data linkage result with Method 3 ... 193

5.5 Representativeness Analysis ... 194

5.5.1 Distributions of key variables ... 195

5.5.2 Summary statistics comparisons ... 201

5.5.3 Bootstrapping analysis ... 204

5.6 Summary 206

Chapter 6 Conclusions 211

6.1 Review of the Research 212

6.2 Potential Contributions 218

6.3 Limitations and Future Research..221

Bibliography 226

Appendices 248

Summary

The past decade has seen a boom and bust in the United States' housing market, and the ensuing financial crisis in the mortgage market has led to an unprecedented amount of housing defaults and foreclosures. While there has been much interest among researchers to understand the triggers and predictors of this crisis, relatively little is known about the channels through which borrower expectations influenced borrower behavior in that period, and about the influence of social networks on borrowers' default decisions. This thesis seeks to investigate how borrower expectations and behaviors affected the crisis, and how these expectations and behaviors were influenced by social networks, in terms of the concentration of foreclosures in surrounding areas.

The first essay is a comparative study of the U.S. single-family and condominium market that investigates the influences of investor behavior and expectations on the U.S. mortgage market between 2003 and 2007. The results show that, over different vintages, condo loans defaulted more often and more quickly than single-family subprime loans, because of the inherently riskier features of condo loans, an assertion supported by loan application data. In addition, condo loans defaulted much earlier than single-family subprime loans, suggesting that foreclosures that resulted from condo loan defaults are associated with higher subsequent defaults in the single-family subprime market, arguably because of the negative spillover effect of foreclosures on neighborhood house prices.

Based on these findings on the influence and significance of borrower behavior in the U.S. mortgage market, there is great necessity to understand the factor that might affect borrower behavior. Therefore, this thesis proceeds to study the impact of neighborhood foreclosure concentration on borrower default behavior in the second essay. Foreclosures are found to induce nearby borrowers to exercise default options more ruthlessly, especially during a market downturn. Besides the damage to the borrowers and lenders directly involved in the default process, foreclosures generate externalities in neighborhood: they induce more borrowers in the surrounding area to default. This circular reaction can continue and lead to foreclosure cascades. Foreclosures can also discourage borrower delinquency, if borrowers take foreclosures as a signal of lenders' reaction to delinquencies, implying that borrowers are strategic in their default decisions.

Multiple datasets are used in the second essay because of the limited coverage of the data in each dataset. Linking multiple data sources together is becoming more common in research. To find out the appropriate way to do data linkage in empirical studies on real estate, a comparative analysis of the different methods in linking multiple mortgage datasets is conducted: propensity score matching, statistical hard matching, and statistical hard matching with machine learning techniques, i.e. Bayes Classifier and decision-tree Classifier. The results show that propensity score matching, although commonly used in real estate empirical research, is not satisfactory for carrying out data linkage; statistical hard matching with machine learning techniques produces better and more reliable linking results.

List of Tables

Table 1.1 Descriptive Statistics: Condominium vs. Single-Family Loans 4

Table 3.1 Summary Statistics of BlackBox Full Sample 74

Table 3.2 Logistic Analysis of Borrower within-two-year Default: Condo vs. Single-Family 86

Table 3.3 Logistic Analysis of within-two-year Default: Condo Loans vs. Subprime Loans 89

Table 3.4 Controlling for Location and Asset Dynamics of the Condo Market...............93

Table 3.5 Controlling for Lender and Contract-term Differences: Evidence using Freddie Mac Data 100

Table 3.6 Investor Channel Analysis: Evidence from Freddie Mac 105

Table 3.7 Within-One-Year Default Analysis 109

Table 3.8 Do Early Condo Defaults Predict Subsequent Single Family Subprime Defaults? 112

Table 3.9 Do Condo Defaults Predict Single Family House Price Growth?..................114

Table 4.1 Number of Loans in Our Sample by Vintage...132

Table 4.2 Performance of Loans in Our Sample..132

Table 4.3 Summary Statistics of Loan and Borrower Characteristics 134

Table 4.4 MLE Estimates of the Cox Proportional Hazard Model...............................145

Table 4.5 Hazard Model Estimates based on Alternative Neighborhood Foreclosure Concentration Measure 147

Table 4.6 Hazard Model Estimates w.r.t. Different Housing Market Regimes..............149

Table 4.7 Asian Borrowers vs. Non-Asian Borrowers...151

Table 4.8 African American Borrowers vs. the Rest of the Population.........................152

Table 4.9 Female vs. Male Borrowers 153

Table 4.10 High FICO vs. Low FICO Neighborhoods 156

Table 4.11 High Income vs. Low Income Neighborhoods...158

Table 4.12 Hazard Model Results for All Subprime Loans in the State of California....160

Table 5.1 Variable list and definitions 175

Table 5.2 The Frequency Distributions of Key Common Variables used in Data Linkage: BBX vs. HMDA 178

Table 5.3 Number of linkages for the three matched groups: 2001-2010......................194

Table 5.4 Summary Statistics of BlackBox: Original vs. Matched Groups 203

Table A1 Summary Statistics of Freddie Mac Full Sample 256

List of Figures

Figure 1.1 Percentage of Homes with a Mortgage in Negative Equity across the United States by County 6

Figure 3.1 Number of condo loans originated in 2001-2007 (in thousands) 63

Figure 3.2 Percentage of condo loans originated in 2001–2007 63

Figure 3.3 Frequency Distribution of Defaults within Two Years of Origination: Condo vs. Single-Family 83

Figure 3.4 R square compassion from OLS regressions: condo and single family by year 96

Figure 4.1 Foreclosure Concentration for LA MSA: for all years 128

Figure 4.2 Foreclosure Concentration for LA MSA: for Year 2003, 2006, 2009 and 2012 129

Figure 5.1 Kernel Density Plots of the BBX variables from the Original and Matched Samples: 2001-2010..196

Figure 5.2 Kernel Density Plots of Borrower Income from the Original and Matched Samples: 2001-2010..199

Figure 5.3 Frequency Distribution of Borrower Information from the Original and Matched Samples: 2001-2010 200

Figure A1 Kernel Density Plots of the original loan amount from the Original and Matched Samples: by loan origination year (2001-2010) 248

Figure A2 Kernel Density Plots of the original LTV ratio from the Original and Matched Samples: by loan origination year (2001-2010) 250

Figure A3 Kernel Density Plots of the FICO score from the Original and Matched Samples: by loan origination year (2001-2010) 252

Figure A4 Kernel Density Plots of borrower income from the Original and Matched Samples: by loan origination year (2001-2010) 254

Chapter 1 Introduction

1.1 Background and Problem Statement

The significant and steady increase in the volume of residential mortgages in the United States has attracted much attention over the past several decades. America has just witnessed one of the longest housing market booms in history, leading to an imbalance in supply and demand. Low mortgage interest rates, low down-payment requirements, various financing alternatives, and relaxation of lending standards lowered the barriers to home ownership.

Nevertheless, credit risk increased because of the explosive growth in mortgage lending between 2000 and 2005, which was followed by the collapse of housing prices in 2007/8. This increase in mortgage lending and credit risk resulted in a substantial surge in residential mortgage delinquencies and a collapse in the values of mortgages. This widespread rise in default rates and the resulting losses in mortgage-backed securities led to a further increase in foreclosures and large decline in house prices, especially in Sun Belt states like Arizona, California, Florida, and Nevada, and Rust Belt states like Michigan and Ohio, [1] marking the start of the market's decline (Lee, 2011). Thus, the rapid pace of foreclosures and house price falls exacerbated the crisis in the housing market, which had been set off by the financial crisis.

[1] Sun Belt states (or sand states) are well-known as bubble states characterized by a relaxed lending market and overbuilding. Rust Belt states have been experiencing a weak economy because of the collapse of the U.S. manufacturing industry.

Introduction

The mortgage crisis and the ensuing uncertainty in the financial markets led to various discussions among academia and practitioners on the possible triggers of this crisis. Some discussions have focused on subprime mortgages (Agarwal et al., 2012), arguing that the high delinquency rate in the residential subprime mortgage market led to severe liquidity shocks and thus the economic depression. Other analysts have discussed the relationship between the growth of securitization and the real estate bubble and crisis (Keys et al., 2010a, 2010b, and Piskorski et al., 2010; Agarwal et al., 2011; An et al., 2011). Some researchers blamed market participants, such as appraisers and mortgage originators, for the risky mortgage products they designed or for mispricing default risks (Agarwal et al., 2012; Ben-David, 2011, 2012).

A less-discussed but fundamental issue is the role of homebuyer and investor expectations. At the peak of the housing boom, home price expectations behaved at abnormal levels; when the housing bust occurred, these expectations fell sharply. The abnormal expectations contributed to the existence of housing investors or speculators, and were highly related to price changes in the housing market (Case et al., 2012). From the demand side, abnormal home price expectations influenced housing speculators to chase short-term trends by speculating on house prices during the boom period and selling them during the market downturn (Bayer, et al., 2011; Chinco and Mayer, 2012; Fu and Qian, 2013). What's more, when prices turned downwards, investors or speculators were more willing to default than other mortgage borrowers, contributing to the rise and fall of the U.S. housing market in the recent crisis (Haughwout et al., 2011). Also, investors are less risk-averse and tend to use unconventional mortgages. The designs of those mortgages, with low interest rates, low down-payment requirements, and relaxed borrower screening criteria, encouraged investors to default (Garmaise, 2013a, 2013b). These effects reciprocally reinforced each other to create a cycle that led to further defaults and contributed to the financial crisis (Campbell, et al., 2011; Mian, et al., 2012). More explanation and evidence of the unique characteristics of investors are given in Section 2.3 of Chapter 2.

However, empirical evidence on the roles of homebuyers' expectations and investor behavior is scarce. This is because of the difficulty of obtaining individual investor data, meaning that most analysts can only study investor behavior using macro-level data. A second reason for the scarcity of individual-level data is the mix of investment and consumption use in the housing market, which means that it is difficult to distinguish investors from consumers. Therefore, due to data limits and identification issues, the influence of the default behavior of investors to the crisis has not been well-studied.

The U.S. condominium loan market provides a unique opportunity to identify and analyze investor behavior. The statistics in Table 1.1 indicate that a significantly higher portion of condominium borrowers are investors. Exotic mortgages are used more often in the condominium market, and less conventional mortgage contracts (such as mortgages requiring low or no documentation, interest-only mortgages and option ARMs) are typically associated with higher-income and self-employed borrowers.[2] Condominium borrowers have higher FICO scores and tend to purchase in more expensive areas. Following Kain and Quigley (1972), Agarwal (2007) analyzed the homeowners of multi-family houses (condominiums) and found that condo owners overestimated the value of their houses by as much as 4.5%, implying a high possibility of them being investors. Therefore, the pattern of defaults in the condominium loan market is likely to be a reflection of the prominent role of expectations in the crisis.

Table 1.1 Descriptive Statistics: Condominium vs. Single-Family Loans

Summary statistics for BlackBox (BBX): from 2003 to 2007

Variable	Total	Condo	Single-Family (SF)	Diff. (Condo-SF)
FICO score	683	699	679	20***
D_Owner occupied	73%	69%	74%	-5%***
D_Option ARM	5%	8%	4%	4%***
D_Low/No doc	35%	41%	34%	7%***
D_Interest only loan	22%	29%	21%	8%***
Log_HPI	5.33	5.36	5.32	0.04***
Sample Size (*1000)	5,000	909	4,091	

Note:

This table presents the summary statistics of the BlackBox Logic (BBX) dataset. Details and explanations of the variables are shown in Table 3.1, Chapter 3.

[2] A low or no documentation loan refers to a finance product offered by a mortgage lender to borrowers who: a) do not qualify for normal loan products or b) do not wish to give up their financial privacy. Borrowers in the first group are often defrauded by brokers who falsify their incomes. In contrast, those in the second group are financially well-off and less likely to be defrauded; however, their incomes are volatile because they are self-employed. Many of the borrowers of low- or no- documentation loans were self-employed (Farris and Richardson 2004). Those who wanted to obtain low or no documentation loans qualified using their high income, liquid assets, good debt-to-income ratio, and a low loan-to-value (LTV) ratio, which is consistent with the characteristics in condominium market.

Introduction

The context described above motivated the first study which examined the role and behaviors of borrowers in the condominium market, especially those who were investors, during the financial crisis. This study contributes to our understanding of the demand-side view by providing evidence that investor behavior, as manifested in the pattern of loan defaults in the U.S. condominium market, plays an important role in explaining mortgage defaults in the crisis. More evidence for this statement is provided in Chapter 3.

In addition to the recent crisis, academics and practitioners have also sought to understand why house owners defaulted on their loans, and the impact of mortgage defaults and foreclosures on the market and individuals. One line of theory for explaining default and subsequent foreclosure is the presence of insufficient equity or negative equity in the property. When a property's value falls below its mortgage value, borrowers may default to maximize their wealth. Such defaults are often called "ruthless defaults" (Foster and Van Order, 1984, 1985). There is a strong relationship between negative equity and defaulting (Quigley and Van Order, 1991; Foster and Van Order, 1984, 1985; Clauretie and Sirmans, 2003). At the end of the first quarter of 2009, 20% to 27% of all homeowners with mortgages were in a situation of negative equity or were "underwater", that is, their debt obligations exceeded their home's market values.[3] Until the first quarter of 2014, although the national negative equity rate has

[3] Deutsche Bank estimated that approximately 14 million U.S. homeowners had negative equity, which was approximately 27% of all homeowners with mortgages at the end of the first quarter of 2009. The real estate website Zillow.com estimated that approximately 20 million homeowners had negative equity at the end of the first quarter 2009. Economy.com estimated that approximately 15 million homeowners had negative equity at the end of the first quarter of 2009 (Weaver and Shen, 2009).

continued to decline since the first quarter of 2012, more than 9.7 million homeowners with a mortgage still remain underwater, as shown in Figure 1.1.

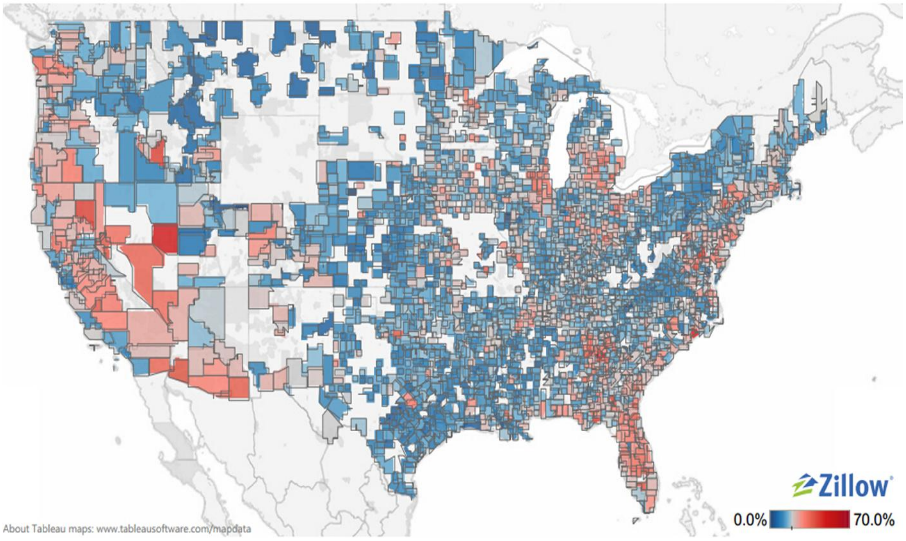


Figure 1.1 Percentage of Homes with a Mortgage in Negative Equity across the United States by County

Source: Zillow Real Estate Research Report March 2014

Note:

This figure presents the percentage of homes with mortgages in negative equity. The colour scale is centred at 18.8%, the national average. Blue counties have fewer underwater homes than the national average, while red counties have more underwater homes.

As a result of the numerous home owners who were in negative equity, many residential defaults and foreclosures were recorded in many parts of the United States. Around 3.2 million of these, an increase of nearly a million since 2007, were identified at some stage (default notices, auction notices, or bank repossessions) of the foreclosure process in 2008 (RealtyTrac, 2009). Cities in the four Sun Belt states accounted for all of the top 20 foreclosure rates in 2009 (RealtyTrac, 2010).

Introduction

The impacts of defaults and thus foreclosures can be devastating on various levels. First, from the perspective of the entire market, high default rates in the residential mortgage market led to severe liquidity shocks at many financial institutions, creating substantial shocks to the U.S. and even global economies. Second, from the perspective of individual participants, defaults and foreclosures led to significant costs and hardships, including the loss of home equity and a potential lack of access to stable credit. Third, from the perspective of surrounding neighborhoods, the rapid increase in mortgage delinquencies and foreclosures had significant negative spillover effects (Lee, 2011). These foreclosures are likely to be spatially concentrated within metropolitan areas, particularly in stressed housing markets in neighborhoods where subprime and other exotic mortgages are more prevalent (Gramlich, 2007; Immergluck, 2008a; Sanders, 2008; Ding and Quercia, 2010). In addition, the increasing concentration of foreclosures and abandoned properties in a neighborhood can result in a rise in violent crime, vandalism and neighborhood deterioration (Baxter and Lauria, 2000; Apgar et al., 2005; Immergluck and Smith, 2006a, 2006b; Kingsley et al., 2009). Property values in these neighborhoods usually decline or stagnate. This price-depressing effect has been widely studied and clearly documented over the last five years (Immergluck and Smith, 2006; Harding, Rosenblatt and Yao, 2009; Daneshvary, et al., 2011; Ding et al., 2011; An and Qi, 2012; Goodstein et al., 2012).

Many communities in the United States have been facing the problems of increasing and concentrated foreclosures for several years. Recently, foreclosure concentration has attracted much attention from the media, home buyers, lenders, economists, researchers and policy makers, because of its role in the housing crisis. However, the mechanisms through which foreclosures influence the decision-making of neighbors, especially their likelihood of and attitudes towards exercising their mortgage default option, have not been well-studied.

Existing studies discuss two simultaneous but contradictory mechanisms that may influence borrowers' default decisions: the information effect and the foreclosure contagion effect. On the one hand, concentrated foreclosures in one's neighborhood can send out a negative signal to nearby borrowers that they are less likely to receive desirable loan modification after defaults, thus discouraging them from exercising the option to default (Guiso et al., 2013; Towe and Lawley, 2013). However, on the other hand, the concentration of foreclosure can induce more defaults due to contagion. Such foreclosure contagion can arise from observational learning[4] or ethical reasons,[5] or as behavioral responses such as herding (Agarwal et al., 2011; Seiler et al., 2012). However, regardless whether the information effect or foreclosure contagion effect dominates, the impact of foreclosure concentration on the attitudes of neighbors towards defaulting on their loan has not been addressed. This topic is further discussed in Section 2.4 of Chapter 2.

[4] Seeing foreclosures in one's neighborhood can cause the borrower to adjust down her property valuation or to strengthen her belief in a declining market, and increase her chance to exercise the default option (Agarwal et al., 2011).

[5] Knowing that many others in the neighborhood have defaulted their mortgage loans might change someone's view that default is immoral or ease the stigma effect of default.

Introduction

Motivated by the background of foreclosure concentration and the problems associated with it, the second research study examines how the display of foreclosure signs in one's neighborhood affected one's likelihood to and final attitude towards exercising her option to enter into mortgage default. By focusing on the Los Angeles-Long Beach-Santa Anna metropolitan statistical area (the Los Angeles MSA), this study provides substantial evidence that the contagion effect dominates the information effect: borrowers are more willing to enter into default when there are many foreclosures in their neighborhood. However, these impacts vary in different regimes and across different groups of borrowers. The details of the methodologies and the evidence are provided in Chapter 4.

For the second study, multiple data sources had to be linked to one another to obtain information on individual borrowers and their loans, because of the limited coverage in each dataset. Various data providers offer a range of datasets, some similar to each other and some unique, for different research purposes. While each of these datasets provides certain information, they often lack useful related information, because of the limitation of their sources or of concerns around confidentiality. Thus, no single dataset has all of the information required for a research project. Given these data limitations, there is a need to link records in two or more separate but related data sets.

Datasets are linked by using variables common to both data sets to identify identical or similar records. This leads to the creation of a new synthetic data set that allows more flexible analysis than would be possible with the two discrete data sets. The main motivation for creating a synthetic data set would be to integrate variables that are never observed together. It is thus important when linking datasets to find the best possible way to match records in the datasets being combined. Database linkage methods generally use both deterministic and probabilistic linking algorithms. Deterministic linkage techniques can be used when both datasets provide record-identifying information that can be matched. However, regulatory and legal restraints on data mean that this approach has limited use. Without common identifiers (such as a residence's unique ID number), a probabilistic linking method is needed to link datasets, and has been studied by academics (Kum and Masterson, 2008; Blanchette et al., 2013).

In recent decades, the most frequently used probabilistic data linkage approach in the field of real estate studies is statistical hard matching which matches the records exactly based on common attributes among the datasets (Haughwout et al., 2009; Reid and Laderman 2009; Ferreira and Gyourko 2011; Ghent et al., 2011; Voicu et al., 2011; Agarwal et al., 2012; Hernandez-Murillo and Sengupta 2012; Pace and Zhu, 2012). Compared to other approaches, statistical hard matching is the easiest to understand and can be used the most directly. However, this method could lead to selection bias resulting from the use of only a few of the observed covariates. Another frequently used method is propensity score matching, which matches records based on the same or similar propensity scores (Rosenbaum and Rubin, 1983; Rubin and Thomas, 1996; Zhou and Lam, 2007; Kum and Masterson, 2008; Fraeman, 2010; Westreich et al., 2010). The propensity score matching method tries to ensure that any differences between the two records from two datasets are not a result of differences on the matching variables, and is ideal for making casual inferences. However, biases, such as regression towards the mean, in propensity score matching may occur if there is limited overlap between the two groups on the matching variables, meaning that the results of the matching exercise may not be representative of the general population. This issue motivated the third study to search for alternative linking method useful in real estate studies.

Recently, machine learning, which can automatically detect intrinsic patterns among the covariates in a dataset and use the uncovered patterns to predict future data or make other kinds of decisions under uncertainty, has been used for linking datasets (Murphy, 2012). Machine learning can identify potential and unobvious patterns in the data that human experts may not be able to, especially when the dataset is huge or has imperfect data quality (Setoguchi et al., 2008; Lee et al., 2010). While the advantages of this approach have been theoretically confirmed in other fields such as computer science, statistics and medicine, it is seldom used in real estate studies, perhaps because of its complexity or a lack of familiarity.

Combining multiple sets of data into a single new one requires a matching procedure that must satisfy a key concern: in the new dataset, the range of values for the measure of interest should be representative of the level of the entire population in the original datasets. Therefore, a matching procedure that preserves at least the marginal distributions of the variables of interest is needed. Also, since the quality of the matching exercise depends on the quality of the data being matched, there is no universal linking procedure that ensures the best quality in all situations exists.

Mortgage data, the main data used in the whole thesis, possess several characteristics that require an effective linking method. First, mortgage data usually track individual loans from certain lenders but with distinct loan information. Therefore, although there might be no unique identifiers among distinct mortgage data, loans from different data but the same group of lenders can be linked, if there are certain common attributes among these data. Second, the mortgage data provide distinct information at certain stages, i.e., loan application, loan origination, and loan termination. Thus, searching a way to link these data at different stages can help provide an overall idea about individual loan performance, probably through application to origination to termination. Third, because of data constraints and different collection methods, there might be few common variables among the mortgage data, thus might not capture the distinct characteristics of the loan records. The fact that the frequently used linking approaches mostly depend on common variables makes it uncertain whether the linking is effective.

Therefore, due to the empirical and technical constraints of the current approaches, and the distinct characteristics of mortgage data, there is a need to compare different approaches to see which approach produces the most representative linking sample and is the most suitable for remedying data linkage issues. Here, two critical mortgage datasets are analyzed and used to compare the different data linkage approaches. The descriptions of the different linking approaches and their differences are presented in Section 2.5 of Chapter 2, and the comparison results and implications are detailed in Chapter 5.

1.2 Objectives and Research Questions

This thesis was separated into three sections: i) a study that analyzed mortgage default behavior in the U.S. condo market and its role as a potential trigger of the financial crisis, ii) a study that examined the effect of foreclosure concentration on borrowers' default decisions in the context of the Los Angeles MSA, and iii) a study that compared various record linkage approaches to guide those facing data linkage problems in real estate research.

The results of these three studies on credit risks, borrower behavior and residential mortgage defaults in the U.S. mortgage market will have various implications for borrowers, lenders, financial institutions, economists, researchers and policy makers. The main objectives of each study are listed here. First, this thesis documents the unique risk pattern among different types of borrowers, especially investors, and their behavior in the U.S. condominium (condo) loan market in the early 2000s. Second, this thesis quantifies the effects of the concentration of foreclosures in certain neighborhoods on the decisions of borrowers to default on their mortgages. Third, the comparison of different ways of linking records will help ascertain the most suitable approach for linking records in mortgage research.

Introduction

The first study, presented in Chapter 3, addresses the default probability of condo loans relative to single-family mortgages, which are more commonly studied, conditional on variables such as characteristics of loans and borrowers and macroeconomic conditions. In addition, competing explanations for the rapid increase in defaults in the condo loan market are examined. These include the unobserved heterogeneity issue, i.e. the presence of unique characteristics in the condo home markets, the lender (supply side) effect, and the borrower (demand side) effect. The following questions are investigated: do condo loans differ from single-family loans with respect to default patterns? If yes, what factor drives the pattern of defaults in the condo loan market? In a neighborhood with condo loan defaults, do condo loan defaults due to risky borrowers have negative spill-overs onto neighboring single-family loans? Or do early condo defaults predict the subsequent default rate in neighboring single family subprime markets? The results show that there is a sharp increase in condo loan defaults relative to single-family loan defaults over the years. The condo loan default rate has also grown at a faster rate, even compared with subprime loans. The unique pattern of defaults in the condo loan market is due neither to the unobserved factors in condominium market compared to the single-family market, nor lender preferences and/or expertise with loans in the condo market and the single-family market. The unique default pattern in condo loan market arose out of the inherently riskier background of condo loan borrowers, compared to single-family loan borrowers. Among all condo loans, investment-purchase condo loans were much more likely to default compared to other condo loans. This relationship was strengthened when the option to default is more profitable (“in the money”). Last, not only do condo loans default earlier compared with single-family loans originating in the same cohort, but these earlier condo loan defaults prompt more defaults in the single-family sector in the same area afterwards.

The results in the first study imply that defaults and thus foreclosures may influence the probability that nearby borrowers will default on their loans. The spillover effects of foreclosures on their surrounding neighborhoods have been well-studied; the effects include house price declines, an increase in violent crime and thefts leading to community instability, an acceleration of racial transition, and broader emotional and physical impacts on individual residents. However, the impact of the concentration of foreclosures on the sensitivity of borrowers to negative equity, or their attitude towards exercising the option to default, has not been fully discussed. It is an open question as to whether the information effect or the contagion effect is more dominant for this issue. This observation motivated the continued analysis of the impacts of foreclosure concentration. The following questions are examined: what is the ultimate impact of foreclosure concentration on borrowers’ attitudes towards mortgage decisions? Or does neighborhood foreclosure concentration have positive or negative impact on borrowers’ decision choices? Is the impact constant or time-varying across different time periods? Do the concentration effects vary among different borrower groups, and among different neighborhoods? Chapter 4 answers these questions by conducting foreclosure intensity measures in two ways. First, it is calculated as the total number of foreclosures in the past two quarters (e.g. for 2009Q1 it is 2008Q4 and

2008Q3) divided by the total number of housing units (in thousands) in each zip code. Second, it is calculated as the total number of foreclosures in the recent four quarters (current quarter plus the past three quarters) divided by the total number of housing units in each zip code (in thousands). The results reveal that, on average, neighborhood foreclosure concentration enhances borrowers' willingness to exercise their option to default in the period that was studied. However, the relative impact of the information effect and the contagion effect differs across different regimes and different borrower groups.

Analyzing borrower behavior and foreclosure concentration in Chapter 3 and Chapter 4 highlighted that each of these datasets provided only some of the information that was needed, because of the limitation of data sources or confidentiality restrictions. Most datasets lacked a significant amount of the information that was needed, meaning that no single source of data had all of the information required for each study. Given these challenges, records in two or more separate but related datasets have to be linked in each study to overcome the limitations of existing data sources. The issue is exacerbated by the absence of shared identifiers in datasets.

However, current linking approaches such as statistical hard matching and propensity score matching have their own limitations when they are used to link multiple datasets. Thus, a more advanced technique from computer science, machine learning, was used in Chapter 5 to link the two mortgage datasets, and the results were compared with other approaches. The following questions are examined: what is the best method to use for linking multiple datasets in real estate studies, especially mortgage studies, when there are no unique identifiers? During the process of data linkage, how can we minimize selection bias and identification errors? Answering these questions extends our understanding of the limitations and potential of different approaches. Thus this study can provide guidance for future real estate researchers when they encounter the probabilistic data linkage.

To answer these questions, a variety of methods on linking the same groups of datasets are tested and compared, including statistical hard matching, statistical hard matching with machine learning techniques (Naïve Bayes Classifier and Decision Tree Classifier), and propensity score matching. Firstly, for statistical hard matching, I use a SAS program to link the selected common attributes of BBX and HMDA data, se well as checking for and eliminating observations with duplicate records. Second, classification algorithms such as Naive Bayes and Decision Tree are applied with statistical hard matching to understand the intrinsic correlations of the key variables, including but not limited to selected common attributes, and produce the learned models for identifying the true matches. Third, with propensity score matching, the BBX and HMDA records with the exact same propensity scores or if they are similar within the same three digits after the decimal point of propensity scores) are regarded as a match; the SAS program is applied to check for and eliminate observations with a duplicate BBX id when there are multiple matches. Across the three approaches, slightly more linked records were obtained when statistical hard matching (Group 1) was used compared to when statistical hard matching with machine learning (Group 2) was used, with the fewest linked records being obtained by the propensity score matching approach (Group 3). Next, in examining the representativeness of the linked samples to the entire population of mortgages, several representativeness analyses were conducted. These included examining the distributions of the key variables (by looking at the kernel density distributions for continuous attributes and frequency plots for categorical attributes), comparing the summary statistics of the matched and original samples, and conducting a bootstrapping analysis on the probability of default based on the key variables from both datasets. The results generally indicate that the statistical hard matching with machine learning approach did a better job in dealing with selection bias and misclassification compared to pure statistical hard matching and propensity score matching. However, the performance of statistical hard matching, while not the best, is generally acceptable when there are no alternatives. Propensity score matching, although well packaged in various programs, should be used more carefully.

1.3 Significance of the Research

The significance of this research project can be seen in its enrichment of existing knowledge in the field, and the practical implications of the findings to the issues faced by practitioners.

This study contributes to the existing literature in at least five ways. First, it is the first to document a strong, robust and economically important default pattern in the much-ignored condominium loan market. Specifically, as the findings of Chapter 3 show, the loan origination growth rate and pattern of default in the condo market are comparable to the subprime mortgage market (Demyanyk and Van Hemert, 2011). Condo borrowers are less likely to have low-quality credit and to default because they could not afford to pay or refinance their mortgages once house prices began declining. In the sample studied, compared to the single family market, condo borrowers have higher FICO scores, use subprime mortgages less frequently, and on average are charged a lower interest rate. These characteristics of condo borrowers suggest that there is a larger proportion of investors in the condo borrower population. Therefore, the condominium loan market provides a unique opportunity to identify and analyze investor behavior. However, studies normally focus on the single-family loan market; the risk patterns and behaviors of condo borrowers are under-studied. This study fills this gap by examining the less-studied condo market to derive some implications for academia and practitioners.

Second, the findings in Chapter 3 also add to our understanding of the economic channels that explain the financial crisis. A large strand of the literature has focused on subprime mortgages and other supply-side factors, such as the role of securitization. On the other hand, recent work (Case et al., 2012; Haughwout et al., 2011) suggests that a less-studied but potentially important factor may have triggered the crisis: homebuyer and especially investor expectations. However, empirical evidence on the roles of homebuyers' expectations and investor behavior is limited, due to the difficulty of obtaining individual-level investor data, or the difficulty in isolating investors from consumers for the purposes of studying them. These data limits and identification issues make it difficult to study how the behavior of investors in terms of their propensity to engage in default contributed to the crisis. The findings in Chapter 3 complement the demand-side view by providing evidence that investor behavior, as manifested in the condo market’s loan default pattern in our context, play an important role in explaining mortgage defaults in the crisis. The results show that loans used for the purchases of condos as investments explained the pattern of defaults observed in the market for condo loans, resulting in the recent crisis. Therefore, this thesis’ study of the characteristics and delinquency probabilities of condo loans from the perspective of borrowers is not only academically meaningful, but also important in explaining what occurred in the recent crisis.

Third, besides the intrinsic triggers of the recent crisis, the decision of mortgage borrowers to default and thus allow their banks to foreclose on their mortgages is an important issue, more so recently with the increasing number of delinquent loans in residential real estate markets in the United States. Understanding why mortgage borrowers decide to default on their loans is also critical for managing the risk of default, and pricing and underwriting mortgages. Traditional studies of the default decision of borrowers focused on their socio-economic status using indicators such as their FICO scores, income constraints, and equity position. Recently, some studies have tried to place borrowers into social networks to understand their default decisions (Gangel et al., 2013; Guiso et al., 2013; Seiler et al., 2013). The study in Chapter 4 follows this line of thought. However, unlike the existing studies that use simulated or survey data, this study uses actual default data. The findings indicate that the behavior of near neighbors strongly influences a borrower's decision to default on his/her mortgage. Knowing that foreclosure concentration affects the decision of borrowers to default signifies that default models should incorporate such network effects to predict the default risks of borrowers.

Fourth, from the data perspective, the analysis on probabilistic data linkage in Chapter 5 is useful in filling in additional or missing information, by adding in extra attributes. With more complete information on population units, more complex research questions can be answered. Linking multiple datasets might help in checking the accuracy and reliability of survey or administratively-collected data, or vice versa. Last, data linkage can enhance data quality by providing more information for people to understand the non-response or non-reported aspects of current datasets.

Fifth, from the linkage approach perspective, the comparative analyses of different linking methods on multiple mortgage datasets improves our understanding of the advantages and potential limits of each method, including their ability to overcome selection bias and misclassification issues. This exercise provides guidance for future real estate studies which also require data linkage when there are no unique identifiers.

This research provides significant policy implications as well. First, the finding that condo borrowers, especially investors, are riskier suggests that lenders need to exercise more scrutiny in their lending practices in the condominium mortgage market. From a public policy point of view, the recent the Dodd-Frank regulations that require lenders to have more “skin-in-the-game” and mandate lower loan-to-value ratios for borrowers are only a partial solution for avoiding a similar crisis in the future. More careful policy to manage the behavior of investors or speculators should be made.

Second, understanding the impact of foreclosure concentration on the decision of borrowers to become delinquent on their loans is also important from a policy perspective. Delinquency is the first step of loan default, and foreclosure is usually the last step. Typically, large numbers of foreclosures follow a wave of delinquencies. The study in Chapter 4 finds that concentrated foreclosures can lead to greater levels of borrower delinquency. While foreclosures are a bad result for borrowers, lenders and investors, the damage was not limited to those parties directly involved in the default process. The foreclosures generated externalities – they induced more borrowers in the neighborhood to default on their mortgage loans. Therefore, during such crises, mortgage defaults can be self-reinforcing in certain neighborhoods: increased delinquencies lead to more foreclosures, and concentrated foreclosures lead to even more delinquencies. This cycle can go on and on and lead to foreclosure cascades. Therefore, it is important for the government and lenders to intervene to stop or reduce foreclosures to break the loop and stop the foreclosure cascade.

Third, the results of Chapter 4 show that the impact of foreclosure concentration on the decision of borrowers to default is not limited to the contagion effect. Sometimes the impact can be on the opposite direction: foreclosures can discourage borrowers from becoming delinquent if borrowers take foreclosures as a signal of how lenders will deal with delinquencies. This information effect is seen to dominate the contagion effect during the market boom. From this perspective, borrowers are strategic in their default decisions. In the future, credit risk modelers should take this game feature of mortgage default into consideration to better understand and estimate mortgage default risk.

1.4 Summary and Organization of the Thesis

The financial crisis of the late-2000s/early-2010s has been accompanied with much discussion of its causes, individual reactions to the crisis and the triggers of this crisis. Among the factors that may have led to it, homebuyer and especially investor expectations are now considered to be key. The impact of those expectations and behaviors on the mortgage market and thus the crisis has been increasingly discussed among academia and practitioners. However, due to the lack of micro-level data and the difficulty in distinguishing between investors and consumers in the housing market, understanding how the default behavior of investors contributed to the financial crisis has been largely unaddressed.

Meanwhile, the decisions by borrowers to default on their loans and foreclose on their homes have been integrated into lender’s decision-making and government’s actions, given the greater awareness of the tremendous financial loss and social instability that results from those decisions. Although understanding how mortgage borrowers make their default decisions is critical to mortgage default risk management, pricing and underwriting, conventional research on borrower decisions focuses mainly on the socio-economic status of mortgage borrowers; the influence of social networks remains an open question.

This thesis focuses on credit risk and borrowers' default behaviors during the 2000s. It first analyses the expectations of homebuyers and especially investors, and provides empirical evidence on the role of borrower default behaviors on the mortgage market and the crisis, using the unique U.S. condominium market as natural experiment. Secondly, it studies the impact of a particular social network - foreclosure concentrations in neighborhoods- on borrowers' final foreclosure decisions in the U.S. mortgage market during the 2000s. This was done to help the government, lenders and related institutions understand the need for timely actions to break the cycle of foreclosures. The need to link multiple datasets for the second study as well as for other studies leads to a discussion of the most appropriate way to deal with data linkage issues in real estate studies, by comparing various approaches used in the real estate field as well as in other fields. This analysis suggests a better linking approach for future studies in real estate.

This thesis is organized as follows. Chapter 2 reviews the relative literature for each of the following chapters. Chapter 3 presents the first essay, entitled "*The Hidden Peril: The Role of the Condo Loan Market in the Recent Financial Crisis*". It examines the important role of borrower default behaviors on the mortgage market and the crisis, using the unique U.S. condominium market as natural experiment. The impact of neighborhood foreclosure concentration on borrower default behavior and the mortgage market is investigated in Chapter 4, entitled "*Foreclosure Concentration and the Exercise of Mortgage Default Options*". Chapter 5 presents the third short essay, titled "*Probabilistic Data Linkage in Real Estate Studies: Applications of Propensity Score Matching and Hard Matching with Machine Learning Techniques*". This chapter explores the relative appropriateness of different approach when faced with the necessity of integrating different datasets and how the quality of linkages can be improved, by analyzing the different linking methods. The final chapter concludes the thesis, highlighting the limitations of the study and offers recommendations for further research.

Chapter 2 Literature Review

2.1 Introduction

The literature review includes four parts. Firstly, I focus on the literature regarding general information on mortgage defaults. The high mortgage default rates in residential mortgage market would lead to cash flow losses to originators in the primary market and also investors in both the primary and second market. The liquidity shocks to these financial institutions create substantial shocks to U.S. and even global economies. In this sense, general literature on mortgage defaults, including the default process, the development history of mortgage defaults and default risks in special markets are critical throughout this thesis.

Secondly, besides the mortgage default literature, my first study is built on the literature of the current financial crisis and its triggers. In reviewing the previous studies about the triggers of the financial crisis, a crucial trigger in the crisis, which is yet largely overlooked in the past and increasing emphasized now in the literature, is borrower behavior, especially investor behavior in mortgage choices. Combining the literature of mortgage defaults and trigger of the financial crisis helps me to test the hypothesis that investor (speculator) behavior plays a significant role in leading to the crisis, and that the condominium loan market is a perfect market for studying the investor behavior.

Thirdly, I briefly review the foreclosure concentration literatures which build up the foundation of Chapter 4. Several studies have comprehensively discussed the impact of foreclosures on surrounding neighborhoods, from the view of house price decline in neighborhoods, rise in violent crime and thefts and thus the instability of the community, acceleration of racial transition, children performance, and emotional and physical impact on people. However, it still remains unclear how seeing foreclosure occurs in one's neighborhood influences someone's likelihood of and attitude towards exercising her mortgage default option to enter into default. Therefore, this group of reviews provides support for the analysis in the fourth chapter, which examines how concentrated foreclosures affect the default decision of mortgage borrowers in the surrounding area.

Fourthly, during the analysis of foreclosure concentration effects, there is a necessity to link multiple datasets, which is also a common requirement in academia. Conventional methods used in real estate studies include statistical hard matching and propensity score matching. The technical constraints of these methods, however, require a more advanced and cleverer method in dealing with the linking issues. Therefore, methods from the field of computer science, i.e. machine learning techniques, as well as the conventional methods are reviewed and compared in this chapter, and these methods are tested and compared in Chapter 5, to see whether the linkage situation is improved by applying the new advanced method.

In this chapter, literature specializing in mortgage defaults is in Section 2.2, followed by a review of studies about recent financial crisis and its potential triggers in Section 2.3. Section 2.4 presents the findings of the literature on the foreclosure neighborhood effects and how concentrated foreclosures affect the surrounding areas and neighbor's decision to default. Then the development and the application of traditional linking methods, i.e., statistical hard matching and propensity score matching, as well as the review of machine learning approaches are presented in Section 2.5. The limitations of each stream of literature will be discussed in each section respectively. Finally, a summary of the literature and the gaps that I am trying to fill is given in Section 2.6.

2.2 General Review on Mortgage Defaults

2.2.1 Mortgage default process

Mortgage defaults are typically analyzed using a dual trigger approach where the first trigger is a shock to the homeowner's income stream. This interruption in cash flow might result from being laid off at work, getting divorced, becoming ill, or even passing away. Once the homeowner is unable to pay, the equity position in the home becomes the second trigger. If the borrower has equity in the property, it makes sense to sell the home, pay all associated fees, and retain the difference. However, if the borrower owes more to the lender than the sale of the home will yield, then there is a possibility that he does not have enough money to pay back the deficiency. But this does not necessarily mean the borrower will default. The homeowner can consider if it is in his best interest to use funds from any number of sources to compensate for the negative equity position (Seiler et al., 2012).[6] If the borrower does pay off the mortgage by borrowing from an outside source, it is

[6] The outside source includes but is not limited to a savings account, borrowing from family or friends, or accessing capital through credit cards, and so forth.

most likely that the new loan will have a higher interest rate. If the homeowner chooses to default on the mortgage (or is otherwise unable to pay off the loan), then he will face severe financial consequences of breaching his mortgage contract. Penalties include a severe reduction in his credit score, difficulty in obtaining future credit, a higher cost when borrowing money in the future, and so forth.

2.2.2 Development of the default theory

Since 1960s, many theoretical and empirical studies have been proposed to explain default risk and default behavior of mortgagors and have developed increasingly mature and comprehensive over time.

A significant stream of literature, beginning in the 1960s and extending through the present, offers the first insights on residential mortgage default risk and addresses default principally from the perspective of the individual mortgage lender. These empirical studies examine the role of loan characteristics and borrower-related factors in default, in order to provide lenders the implications for predicting borrower default probabilities. The early works of Jung (1962), Page (1964), and von Furstenberg (1969), among others, evaluate the relations between mortgage risk and characteristics of the mortgage loan, including the loan-to-value ratio, interest rate, and mortgage term. Subsequent research extends this analysis of mortgage risk to include a series of borrower (von Furstenberg, 1969; Herzog and Earley, 1970; Sandor and Sosin, 1975) and property (von Furstenberg and Green, 1974) characteristics. However, during that period, no attempt was made to provide a theoretical basis for borrower behavior at the time of default.

Since the late 1970s, a lot of studies seek to explain the behavior of individual households through structured models. Rooted in the economic theory of consumer behavior, such studies model the behavior of individual households that, in the course of maximizing their utility (and net wealth) over time, rationally decide whether it is in their best interest to continue making payments on their mortgage loans. Jackson and Kasserman (1980) are the first to support the optimization model of consumer choice in the analysis of default decisions, followed by Campbell and Dietrich (1983) which work on the significance of net equity in the borrower's decision to default.

The evaluation and pricing of default have been frequently discussed since the last two decades. Beginning with Asay's seminal effort in 1978, the Black-Scholes (Black and Scholes, 1972) option pricing model has been applied to the pricing of mortgages and their derivative securities. Default is treated as a put option, allowing the borrower to sell the house to the lender for the value of the mortgage at the beginning of each payment period (Foster and Van Order, 1984).[7] In assessing whether or not to exercise the option, borrowers consider the market value of the mortgage and the equity they have in the home, which is a crude measure of the extent to which the put option is "in the money" (Quigley and Van Order, 1991). Early research about the default risks of residential mortgage and

[7] Similarly, prepayment can be viewed as a call option, allowing the borrower to exchange a sum of money for the mortgage instrument.

economic behavior of mortgage holders employs the standard contingent claims approach to mortgage pricing. The contingent claims models are developed by Black and Scholes (1973), Merton (1973), Cox, Ingersoll, and Ross (1985), and others. These models provide a coherent motivation for borrower behavior, inferring that default and prepayment are options to put and call the contract respectively, as a function of loan attributes such as loan-to-value ratio (LTV) and debt-service coverage ratio (DCR) (Dunn and McConnell, 1981; Buser and Hendershott, 1984; Brennan and Schwartz, 1985; Kau et al., 1995; Harding, 1994; Quigley and Van Order, 1995). Hendershott and Van Order (1987) and Kau and Keenan (1995) have reviewed much of the literature related to mortgage pricing.

Most studies employing option models help explain merely one of default and prepayment behavior: some only consider option-based prepayment models (Findley and Capozza, 1977; Green and Shoven, 1986; Schwartz and Torous, 1989; Quigley and Van Order, 1990), while others only applied option models to price default risk (Cunningham and Hendershott, 1984; Epperson et al., 1985; Foster and Van Order, 1984; Quercia and Stegman, 1992; Quigley and Van Order, 1995; Vandell, 1993). A common limitation of the above studies is the failure to consider default and prepayment simultaneously and interactively: the jointness of the prepayment and default options is crucial in explaining behavior. In addition, these studies do not take into account the effects of contemporaneous cash flow conditions on put and call risks.

Literature Review

A group of papers by Titman and Torous (1989), Kau et al. (1992, 1995) and Kau and Keenan (1996) provide theoretical models which emphasized the significance of the jointly considering prepayment and default options. A homeowner who exercises the default option at current period gives up the option to default in the future, but at the same time he/she automatically gives up the option to prepay the mortgage. Foster and Van Order (1985) estimate simultaneous models of default and prepayment using data on large pools of FHA loans, and Schwartz and Torous (1993) estimate the joint hazard using a Poisson regression approach and aggregate data. Deng et al. (1996) and Deng (1997) are the first to analyse residential mortgage prepayment and default behavior using micro data on the joint choices of individuals. More importantly, Deng et al. (2000) present a unified model of the competing risks of mortgage termination by prepayment and default, considering the two hazards as dependent competing risks which are estimated jointly. This work also accounts for the unobserved heterogeneity among borrowers, and estimates the unobserved heterogeneity simultaneously with the parameters and baseline hazards associated with prepayment and default functions. From the perspective of empirical matter, Deng and Quigley (2002) consider that mortgage holders do not behave as ruthlessly as the theory predicts. They develop an option-based empirical model to analyze the behavior of irrational or bounded rational “woodheads”: there exist a group of borrowers who forego substantial savings on mortgage payments through refinancing but prepay the mortgage instead. Their results show that, the unobserved heterogeneity is due in part to the non-optimizing behavior, which is the behavior of "woodheads".

Turning to default decisions, option theory predicts that negative equity is the most important variable determining the optimality of default. If there is negative equity in the house, the homeowner can exercise the put option by default to maximize his wealth. However, after years of development in mortgage theoretical model and empirical analyses, more and more studies find that option value is far from enough to explain borrower choices to default – many borrowers do not default although their houses have substantial negative equity (Vandell, 1995; Cauley, 1996; Archer et al., 1996; Clapp et al., 2001; Pavlov, 2001; Deng et al., 2005).[8]

Ambrose et al. (1997) state the significance of transaction costs. Deng et al. (1996, 2000) show that "trigger events" (i.e., shocks to an equilibrium) such as unemployment and divorce are important to the borrower's default decision. Vandell (1995) argues for similar trigger events or shocks, which are crucial to default decision. Archer et al. (1996) summarize mobility driven factors related to mortgage termination into two broad categories: the location decision factors and the response to housing disequilibrium factors. Employment opportunity is the most important location-driven mobility factor; people usually move because of job relocation. Pavlov (2001) finds that the local unemployment rate is positively

[8] For example, Cauley (1996) reports that there was little increase of default rate even though up to 44 percent of homes purchased between 1989 and 1991 in Los Angeles County had negative equity in 1995.

related to move because there might be more attractive opportunities outside the local area. Besides employment, other factors like climate and health are also important location-driven mobility factors. Pavlov (2001) and Deng et al. (2005) also argue that default is primarily driven by the optimality of a move in the presence of negative equity. Moderating variables such as years in the current home or proxies for transaction costs are considered by these studies. There is increasing consensus that household mobility factors are also crucial for default.

Credit risk, the risk of financial loss due to an unexpected deterioration of counterparty credit quality, has doubtless been brought into sharp focus over recent years, but it has also played a significant role in the majority of financial crises prior to this time. FICO score[9] is widely accepted by the lenders as observable information for credit evaluation to capture the risks of mortgagors. Borrowers' credit history information used in FICO determinants includes delinquency (late payments), the amount of time that credit has been established, length of residence, and negative credit records (e.g., default, personal bankruptcies). When lenders use risk-based pricing to incorporate the credit history information into their mortgage pricing, borrowers with credit scores are

[9] FICO risk score is a kind of credit scoring method developed by Fair Isaac&Co. and universal in the residential mortgage field with a range of 300 to 850. There are three largest credit bureaus issuing borrowers credit report and FICO scores including Experian, Transunion and Equifax. Strictly speaking, the lenders actually differ on grades given same FICO borrowers scores. The method of calculating a credit score is to attempt to condense a borrowers' credit history into a single number. The Federal Trade Commission has ruled this to be acceptable. Of course, they are also other credit scores which would try to measure properly the credit history of the borrowers, such as NextGen, VantageScore, and CE score in United States. It is noted that credit score is required to capture the credit history factors, not other discriminate and predatory factors .For example, in American, the Federal Reserve Board's Regulation B (implementing the Equal Credit Opportunity Act), expressly prohibits a credit scoring system considering "prohibited bases" such as race, skin color, religion, national origin, sex, and marital status. Source: http://en.wikipedia.org/wiki/Credit_score_(United_States).

assigned with different credit spreads. Automatic underwriting reduces the operating costs for originating and evaluating individual’s mortgage default risks. Studies on mortgage largely have been concentrated on the role of borrower credit risk and credit constraint in the analysis of mortgage origination and performance (e.g., Gabriel and Rosenthal, 1991; Canner et al., 1994; Bradley et al., 1995; Avery et al., 1996; Goering and Wienk, 1996; Munnell et al., 1996; Ambrose et al., 1997; Berkovec et al., 1998; Ondrich et al., 2000; Ambrose et al., 2001; Ambrose and Sanders, 2005).

2.2.3 Default risks in condominium market

Through the development of the default theories, it is shown that most of the mortgage studies in developed countries such as U.S. focus on single-family loans, those of which are based on the building occupied by just one household or family, and made of just one dwelling unit or suite. The reason is that single-family loans historically and currently account for the largest proportion of housing in U.S., among others. There are very limited studies focusing on the condominium loan market. However, since the late 1960s, the fast growth and popularity of condominium sales have made condominium market critical in both real life and academia. Condominiums are less expensive to purchase and require less maintenance than traditional detached single-family dwellings, yet they offer the same tax benefits as home ownership. The desire of homeowners to be close to cities and the scarcity of land in urban areas also encourage the use of condo housing. Additionally, home buyers often seek properties that include recreational facilities in recent times. Finally, the home buying market is changing towards condominium market because single, divorced, childless, elderly, and geographically mobile consumers keep entering the home-buying market and find that condominiums meet their special housing needs. This group of home buyers is distinct from traditional single-family home buyers, with respect to their characteristics and attitude to credit risks. Given the increasing amount of condo loans and distinct condo borrower characteristics, it is necessary to study whether the condominium loan market has different risk pattern and borrower behaviors than single-family loan market in the U.S.

The mortgage default studies have developed over time and are obtaining greater attention since the tremendous crisis that the U.S. housing market experienced in the last decade.

2.3 Recent Financial Crisis: Triggers

During the past decade, the U.S. housing market experienced two interrelated events. First, it is widely believed that the U.S. experienced a housing market bubble in the early 2000s and that this bubble burst in 2007. Second, during this same period, the use of unconventional mortgage products escalated. Individual mortgage default has received much more attention after the unfolding of turmoil from the last quarter of 2007. Many commentators have looked for the triggers of this crisis.

2.3.1 Innovation in mortgage products

Some have pointed to the innovation in mortgage products (e.g., subprime mortgages).[10] The high default ratio in residential subprime mortgage market has caused liquidity shocks, which subsequently lead to economic depression in U.S. and other countries globally. Regulators, economists, policy makers, politicians, government, agencies, research, speculators, and bankers all look into the melt-down of the "subprime" mortgages, which were mostly issued to low-income, minority borrowers. They attempt to find explanations for the high mortgage defaults, besides sharp housing downturn. They share the view that default is not only a pure financial event triggered by stochastic macro-condition (e.g., housing price and interest rate), but default behavior varies by individuals.

2.3.2 Fast growth of securitization

Some other literature pointed to the remarkable growth of securitization in recent years as a major contributor to the rise of the real estate bubble and the ensuing crisis (Keys et al., 2010a, 2010b, and Piskorski et al., 2010; Agarwal et al., 2011; An et al., 2011). Part of the argument is that securitization has created additional layers of adverse selection and moral hazard problems in loan origination and servicing, which in turn led to lax underwriting, as well as higher default rates (Keys et al., 2010a; Agarwal et al., 2012).

[10] These products were designed to help borrowers in markets expecting significant price appreciation. However, they were often marketed to borrowers with relatively poor credit histories as well. As a result, these mortgages are often referred to as subprime mortgages, since they did not meet the underwriting criteria of the housing government-sponsored enterprises (Agarwal et al. 2012).

2.3.3 The wisdom of market players

Other have questioned the wisdom of the lenders and investors who invested in mortgages backed by overvalued assets (Agarwal et al., 2012; Ben-David, 2011, 2012), and underpriced default risks (An et al., 2012). Specifically, these papers discuss the role of professional appraisers and mortgage originators, and the stability of the models these professionals rely upon for pricing the mortgage default risks. They argue that combining the poor data input, unstable model and human error, the mortgage originators steered borrowers to riskier products during the period leading up to the recent crisis.

2.3.4 Borrower (investor) behavior

Another significant factor in the crisis, which is largely overlooked in the past and increasing emphasized now in the literature, is borrower behavior in mortgage choices. Mortgage choice mostly studies borrowers' self-selection in asymmetric information framework. In these studies, borrowers with different exogenous default risk (measured by exogenous movability, or probability of income changing in two-period models) are suggested to self-select into different mortgages. Firstly, borrowers with exogenous default risk profile self-select into different mortgages with LTV and coupon-points combination. For example, high risk borrowers (measured by exogenous high default costs) will self-select into high loan-to-value ratio; while high risk borrowers (thus low default costs) self-select into high loan-to-value (Chari and Jagannathan, 1989; Brueckner 1994; LeRoy 1996; Stanton and Wallace, 1998; Brueckner 2000; Harrison et al., 2004; Chang and Yavas 2009). Secondly, borrowers with different exogenous risk factors (e.g., socially moving incentive and impatience) prefer different mortgage type (Brueckner 1992; Brueckner 1993; Coulibaly and Li 2009; Dhillon et al., 1987; Follain 1990; Hendershott et al., 1997; Mori et al. 2010; Posey and Yavas 2001; Sa-Aadu and Sirmans 1995). For example, borrowers with low credit scores and high loan-to-value preference are also more likely to choose subprime hybrid adjustable rate mortgages; borrowers with high credit scores and low loan-to-value preference choose prime or subprime fixed rates mortgages (Mayer et al., 2009). The relation between the borrowers' observable characteristics (e.g., LTV, mortgage size) and their unobservable risk (e.g., real income, creditworthiness, borrowers' initial wealth, preference to house consumption and default tendency) contributes to the potential cause of the mortgage default behavior and thus the tremendous crisis.

Among borrowers of mortgages, investors play especially important role in the crisis. Investors in the housing market are observed to possess several characteristics that may result in higher risks in default. First, from the labor market perspective, investors are usually self-employed people, who work for themselves instead of an employer and draw income from a trade or business that they operate. Compared with those taking regular monthly (or annually) salaries, the self-employed have instable income and thus probably are unable or not willing to provide full financial statements or taxation returns to verify their current income. In addition, self-employed people are more unstable in job status: they have jobs today but may lose their jobs in the future. They therefore may under-predict their employment risk, overvalue their houses and are thus more likely to default (Agarwal, 2007; Agarwal, et al., 2005).

Second, from the housing market perspective, probably due to their relatively high current income and investment returns from the market, or due to the high expectation of their future income, investors tend to over-predict the house price appreciation in the future, which result in higher default risks when the housing market collapses. The over prediction of their income and the house price appreciation may also lead to default in the future if they cannot afford the mortgage payment, especially when the housing market collapses.

While much of the existing literature focuses on the innovation in mortgage products (e.g., subprime mortgages), or securitization and the associated agency problems in recent years as a major contributor to the rise of the real estate bubble and the ensuing crisis, some recent studies (Haughwout et al., 2011; Case et al., 2012) suggest the root cause of the recent housing crisis in the U.S. can be attributed to a previously less studied, but potentially more fundamental factor—the homebuyer and especially investor expectations. Based on a survey sample from four U.S. cities, Case et al. (2012) report that home price expectations, which reached abnormal levels relative to the mortgage rate at the peak of the boom and declined sharply since, were highly correlated to the price movements of the housing market. Haughwout et al. (2011) hypothesize that real estate "investors"—borrowers who use financial leverage in the form of mortgage credit to purchase multiple residential properties—played a previously unrecognized, but very important, role to fuel the housing boom and exacerbate the housing bust. Specifically, when prices turned down, they defaulted in large volumes and thereby contributed importantly to the intensity of the housing cycle's downward leg. Barlevy and Fisher (2011) show that speculators are more likely to choose exotic mortgages and more likely to default. Amromin et al. (2011) find that high credit worth households chose complex mortgage products leading up to the crisis and they defaulted more. Case-Shiller hypothesis about the great influence of house price expectation and speculative behavior is well-developed and commonly accepted, consistent with some other studies using macro-level data to support this argument. A few studies are based on micro-level data which show that real estate investors chase price trends, push the prices away from the fundamental level, and are very sensitive to negative market shocks (Fu and Qian, 2012; Fu et al., 2012).

However, empirical evidence on the roles of homebuyers' expectation and investors' behavior is scarce. Firstly, most of the analyses are based on macro-level data, due to the difficulty in obtaining individual investor information. Secondly, housing market, which both investment and consumption behavior exist, is different from stock market. The stock market trades everyday while the housing market has less frequent transactions; stocks traded in the stock market have only investment use, while houses are transacted for either investment or consumption use or the combination of both; in the stock market, people can exit their equity positions quickly and almost without cost, while in the housing market the transaction costs for exiting the asset positions are quite large (Case and Shiller, 1988). Therefore, in the housing market, it is quite difficult to distinguish investment and consumption purpose from the purchase behavior due to the heterogeneity of individuals.[11] Due to the above data constraints and identification issues of investors from consumers in the housing market, it still remains an open question on how the default behavior of investors contributed to the crisis.

The condominium loan market, given the characteristics, provides a unique opportunity to identify and analyze the investor behavior. Using U.S. condominium market as a perfect experiment, we contribute to the existing literature by providing a unique angel to identify the investors from consumers to test the Case-Shiller hypothesis about the influences of investors (speculators) behavior and their expectations on mortgage market.

2.4 Neighborhood Effects of Foreclosure and Foreclosure Concentration

As the national mortgage crisis has worsened in late 2000s, an increasing number of communities are experiencing declining housing prices and rapidly increasing foreclosures. Foreclosures not only hurt those who are losing their homes to foreclosure, but also harm neighbors by reducing the value of nearby properties and in turn, reducing local governments' tax bases, thus calling for the

[11] Some studies regard household as speculator if he purchases more than one property besides the one he actually lives in. However, this is doubtful in real cases. For example, the buyer buys a property that he himself will not live in, but this property is bought for his mother-in-law. Another example is that the property is bought in the vacation place such as Hawaii; his family only goes there for a month every summer. In these two examples, the second house is not for investment use. Therefore, using this criterion to be related to investors lacks support.

government intervention. The extent to which foreclosures do in fact drive down neighboring property values, and how those impacts vary according to neighborhood characteristics and local housing markets, have been highly debated among researchers, and also policymakers as they struggle to address the rising tide of foreclosures throughout the country (Schuetz et al., 2008). There is also a group of literature discussing the contagion effect of foreclosures (Schuetz, et al., 2008; Harding, et al., 2009). However, how concentrated foreclosures affect the default decision of mortgage borrowers in the surrounding area still remain an open question.

2.4.1 General Literature on Neighborhood Foreclosure effects

The impact of foreclosures on the individual or institution holding the failed mortgage has been broadly and comprehensively studied (Kau and Keenan, 1995; Capone, 2001). In recent decades, the neighborhood externalities for foreclosures have been gaining much attention, both in academia and in other fields (Leonard and Murdoch, 2009).[12] There are some empirical studies attempting to quantify the effect of foreclosures on surrounding neighborhoods. Immergluck and Smith (2006) attempt to estimate the effects of foreclosures of one- to four-family homes on the property values of surrounding one- to four-family homes in Chicago. Following this earliest and most frequent city study, Leonard and Murdoch (2009) and Lin et al. (2009) also report that the presence of foreclosed properties is associated with lower sales prices for nearby non-distressed properties. Rogers

[12] For instance, in a May 5, 2008 speech, Federal Reserve Board Chairman, Ben Bernanke, stated ‘‘High rates of foreclosure can have substantial spillover effects on the housing market, the financial market and the broader economy”.

and Winter (2009) state that the rise in foreclosures will result in declines in the sales value of neighboring properties, which, in turn, will lead to an extension of the housing crisis.

There are several possible mechanisms through which foreclosures might have a negative impact on the values of the nearby properties. The first is through a negative visual externality: property owners who receive foreclosure notices may be less likely to maintain or upgrade their properties, either because they have less incentive to maintain property they may lose or because they cannot afford regular maintenance. Properties may start to show visible signs of neglect, which may make the surrounding homes less desirable. The second mechanism, social interaction, is explained by Ioannides (2003) that individuals' valuations of their own homes are influenced by those of their immediate neighbors. In consequence, a decrease in value of a nearby foreclosed property can result in lower seller reservation prices and lower sales prices for nearby non-distressed properties. Foreclosed properties also increase the supply of homes and the sellers of foreclosed properties are highly motived to sell quickly, thus affecting the price of "comparables" used to estimate neighboring property values and putting down the ward pressure on local prices (Lin at al., 2009; Harding et al., 2009). Third, although the motivation to sell the foreclosed properties is strong, after completion of foreclosure proceedings and eviction of the delinquent borrower, the property may still remain unsold and vacant, and thus suffer further physical decline. Vacant properties are likely to depress surrounding property values since they contribute to neighborhood blight, may attract vandalism and crime, and more generally signal that the neighborhood is not stable, thus further influence the neighborhood property value. Finally, distressed properties sold either at foreclosure auctions or pre-foreclosure sales may be more likely to be sold to investors and become renter-occupied, which may lead to lower levels of maintenance even after the properties are re-occupied, thus driving down the values of the properties.

Besides the research on the impact of foreclosures on housing prices reviewed above, several studies have examined the effects of foreclosures on other neighborhood outcomes. Some studies argue that the aftershock of foreclosure goes beyond just homeowners but also expands to towns and neighborhoods as a whole. Cities with high foreclosure rates often witness more crime and thefts with abandoned houses being broken in to, garbage collecting on lawns, and an increase in prostitution.[13] Immergluck and Smith (2006a) use a cross-sectional methodology to examine the effects of single-family foreclosures on crime rates in Chicago, and conclude that foreclosures increase violent crime but not property crime. A set of related studies find that foreclosures accelerated racial transition in New Orleans by depressing housing prices and creating opportunities for lower-income black households to move into formerly white-occupied homes (Lauria, 1998; Baxter and Lauria, 1999; and Baxter and Lauria, 2000). They also observe that higher foreclosure rates were correlated with higher vacancy rates and lower proportions of owner-occupied housing. Apgar et al. (2005) estimate that in the City of Chicago, foreclosures impose substantial costs

[13] Associated Press. "Sharp Rise in Foreclosures as Banks Move in - Business - Real Estate – Msnbc.com." NBC News, 13 Oct. 2011.

upon the municipal government, thereby attract criminal activity or squatters, require physical maintenance and/or incur structural damage from fire or abandonment. Another significant impact from increased foreclosure rates mentioned in the literature is the effect it has on school mobility of children, and thus the potential academic performance for children (Been et al., 2011). Foreclosures also have an emotional and physical impact on people. In one particular study of 250 recruited participants who had experienced foreclosure, 36.7% met screening criteria for major depression (Pollack and Lynch, 2009).

Differences in state laws may shape the neighborhood impacts of foreclosures as well. Differences in foreclosure laws can influence the length of time between initial foreclosure filing and the completed foreclosure. For instance, judicial process states such as New York and Illinois have foreclosure proceedings lasting for a year or more, while in Texas most foreclosures are non-judicial and may be resolved in as little as three months (Bergman, 1996; Nelson and Whitman, 2004). The distinctions in foreclosure process and local housing market conditions imply that even studies using comparable data and methods may reach different conclusions when applied to different parts of the country. Moreover, most of these studies have obtained data on foreclosure filings from different sources, so it is unclear whether even the count of foreclosures is truly comparable across studies. This could lead to problems such as confounding the effects of mortgage-related foreclosures with those of tax liens, or simply an inaccurate count of the number of foreclosures within the time-distance intervals.

2.4.2 Impact of Concentrated Foreclosures on borrowers' default decision

The above studies comprehensively discuss the impact of foreclosure concentration on surrounding neighborhoods, from the perspective of house price decline in neighborhoods, rise in violent crime and thefts and thus the instability of the community, acceleration of racial transition, children performance, and emotional and physical impact on people. Nevertheless, it still remains unclear how witnessing foreclosure signs in one's neighborhood influences someone's probability of and attitude towards exercising her mortgage default option to enter into default.

As social animals, humans knowingly or otherwise look to their peers before reaching financially life-altering choices. As such, the impacts of concentrated foreclosures in one's neighborhood on his default decision are necessary to be studied. On the one hand, from a game-theoretic perspective, concentrated foreclosures in one's neighborhood can discourage the borrower's exercise of default option, by transmitting information to neighbors (Guiso et al., 2013; Towe and Lawley, 2013). This is because intense foreclosures in a neighborhood can send out a signal to nearby borrowers that should they choose to default they are likely to be similarly foreclosed instead of receiving a favorable loan modification. This information effect is similar to that discussed by Riddiough and Wyatt (1994) and Guiso et al. (2013) where borrower's strategic default decision depends on her belief of what the lender's reaction would be: foreclosing loans can prevent other borrowers in the market from strategically defaulting as they perceive the foreclosing banks as “tough” and not so willing to renegotiate.

However, on the other hand, concentration of foreclosure can induce more defaults due to contagion. Such foreclosure contagion can arise from observational learning: seeing foreclosures in one’s neighborhood can cause the borrower to adjust down her property valuation or to strengthen her belief of a declining market, and thus increase her chance of exercising the default option (Agarwal et al., 2011). Foreclosure contagion can also arise from ethical reasons: knowing that many others in the neighborhood have defaulted their mortgage loans might change someone’s view that default is immoral or ease the stigma effect of default. In addition, it can arise from behavioral responses such as herding (Seiler et al., 2012). If this moral hazard problem is allowed to continue, the global recession currently experienced could become much more severe moving forward (Seiler et al., 2013). Foreclosure contagion is suspected of exacerbating the housing crises during the Great Depression and the recent financial crisis (Campbell, 2013).

With these existing studies, the impact of neighborhood foreclosure concentration on individual borrower’s delinquency probability is increasingly emphasized and studied. However, the impact of foreclosure concentration on the borrower’s sensitivity to negative equity, i.e., to a certain extent the changing attitude of borrowers towards default option exercise, has not been fully discussed. Thus whether the information effect or foreclosure contagion effect dominates neighborhood foreclosure concentration impact on nearby borrowers’ delinquency decision is an open question.

2.5 Probabilistic Data Linkage Methodology

In this section, the original work and categories of data linkage which is the foundation of our analysis is first briefly reviewed. I then review the literature on two commonly used approaches when dealing with probabilistic data linkage: hard matching and propensity score matching, especially in the field of real estate finance and economics. This group of reviews helps us understand and compare the advantages of hard matching and propensity score matching among multiple datasets when there is no unique identifier, and also the potential problems associated with these linking methods. These discussions also motivate me to apply these methods more carefully and more creatively. The literature on machine learning techniques is further reviewed, which is well developed and commonly used in the field of computer science. The advantages of machine learning techniques in dealing with selection bias encourage us to apply this approach to our matching mechanism.

2.5.1 General literature on Data linkage

Data linkage, or statistical matching, is by now a widely used technique in producing empirical studies (Kum and Masterson, 2008). The method is originally applied in many observational studies in medical literature, where patients from different database need to be linked together (Rosenbaum and Rubin, 1983; Rubin and Thomas, 1992, 1996; Little and Rubin, 2000).

Data linkage is deterministic if a unique identifier or key of the entity of interest is available in the record fields of all of the data sources to be linked, and it is probabilistic if a unique key is not available so that not all units can be unambiguously identified (Steiner and Cook, 2013). The former situation is straightforward but not frequently occurred, while the latter one, which is also called probabilistic data linkage, is more complicated but also more frequently encountered. Rässler (2002) gives an example of the probabilistic data linkage where researchers are interested in the association between television viewing and purchasing behavior but lack data from a single source panel covering information on both behaviors. Thus, the idea is to combine data from an independent television and consumer panel by matching on similar subjects. For each unit in the consumer panel, the matching task consists of finding a corresponding subject that is identical or at least very similar on the shared covariates. Such matching of subjects is equivalent to imputing missing covariates on the television viewing behavior.

Probabilistic data linkage approach is becoming a standard in social science research (Radner, 1981; Greenwood, 1983, 1987; Wolff, 2000; Brodaty et al., 2001; Wagner, 2001; Rässler, 2002; Keister, 2000, 2003). Specifically, in the field of real estate studies in recent decades, the most frequently used probabilistic data linkage approach is statistical hard matching (Haughwout et al., 2009; Reid and Laderman 2009; Ferreira and Gyourko, 2011; Ghent et al., 2011; Voicu et al., 2011; Agarwal et al., 2012; Hernandez-Murillo and Sengupta, 2012; Pace and Zhu, 2012). The differences among these studies majorly lie in how they choose linking criteria and how they deal with the potential bias from the multiple matches and non-matches.

Some literature relies on random selection approach to deal with multiple matches during each matching step. Ferreira and Gyourko (2011) link DataQuick transactions data with individual loan information from HMDA, by conducting two-step record linkage.[14] From the multiple matches after each step, for the records with the same identifier, only one is randomly selected to be the true match. Under Ferreira and Gyourko (2011)'s matching algorithm, 60 percent of their total matches are claimed "high quality" match. However, the matching is doubtful due to the accuracy of random selection for the true match. Although one can always use as detailed a random selection mechanism as possible, there is still a huge probability that the random selection procedure misclassify actual match as "unmatch", and actual unmatch as "match", due to the limited matching criteria. This misclassification problem may thus cause the unreliable analyses afterwards.

Some other studies try to separate the matching into several detailed steps, repeatedly dealing with multiple matches in each step. Voicu et al. (2011) match the LoanPerformance (LP) and HMDA loans in two steps with detailed matching algorithm.[15] Using this matching algorithm, they manage to link nearly 15 percent

[14] In the first matching step, each transaction in DataQuick is matched to a loan in HMDA by four criteria (year, Census tract, the lender name, and the exact loan amount). In cases where there are multiple matches, one of them is randomly assigned as being a true match while the rest are considered unmatched. In the next step, unmatched observations in DataQuick not being recognized in the first step are then merged to those in HMDA using relaxed criteria: only year, Census tract and exact loan amount; still for multiple matches, one is randomly assigned as true match.

[15] In the first step, they link the two datasets based on six "mandatory" criteria. The six "mandatory" criteria is: (1) the HMDA action year matches the LP origination year; (2) the LP

of the LP loans to HMDA data. Similar to this algorithm, Agarwal et al. (2012), Agarwal et al. (2012), and Pace and Zhu (2012) also match LP to HMDA loans and further study the potential influences of socioeconomic and demographic information on the borrower and lender differences on subprime foreclosure outcomes. Compared with other algorithms, this group of matching algorithm more carefully deals with the multiple matches, which to some extent reduces misclassification bias. However, this approach also faces potential selection bias issues, since it only accounts for observed confounders; those unobserved covariates which cannot be used in the matching procedure are not considered. Therefore, the matching quality is not ideal if there are only a few commonly observed covariates among multiple datasets.

There is also a line of research which does not deal with multiple matches. Haughwout et al. (2009) also link the performance and terms of the loans from LP data with HMDA data, with only one-to-one matches are considered.[16] Ghent et al. (2011) and Hernandez-Murillo and Sengupta (2012) also follow the Haughwout et al. (2009)'s matching algorithm to combine loan-level data with

loan number is contained in the HMDA application number; (3) the Federal Information Process Standard (FIPS) states codes match; (4) the loan amounts match; (5) the lien status matches; and (6) the LP origination date is later than the HMDA application date. In the second step, they use six additional criteria to select one HMDA match for each LP loan, the one satisfying most of these additional criteria. The additional criteria are: (1) occupancy; (2) loan purpose; (3) loan type; (4) originator name; (5) date (if LoanPerformance origination date is within 30 days of the HMDA action date); and (6) zip code (based on identifying zip codes associated with the census tract for the HMDA loans).

[16] They match LP into HMDA in six stages. In the first stage, LP loans are matched to HMDA loans with the same first 4 digits of the loan's zip code, same purpose, occupancy, lien status, origination time, and loan amount within $1,000 difference. After Stage 1 all loans other than one-to-one matches are put into the next five stages, while some criteria are relaxed or tightened (e.g. zip code is matched to 5-digits or origination amount must be exactly the same). Finally after the sixth stage, all one-to-one matches are aggregated into a dataset to be the final sample, while the rest are considered unmatched.

individual- and neighborhood-level data. This matching approach, however, does not pay sufficient attention on the remaining records after aggregating all one-to-one matches, which may again lead to selection bias and misclassification issues.

These studies, consistent with broader economics and finance literature, usually reply on econometrics methods to hard matching multiple datasets. Hard matching can perform well when dealing with a small number of observations; however, this approach is limited when facing a large number of covariates. Also, hard matching does not account for selection bias resulting from limited observed covariates and probably abundant unobserved covariates. What's more, these studies simply omit that hard matching process may result in some errors and incompleteness into the resulting records. As a result, two indeed unmatched entities from two datasets may give rise to identical records (either due to errors or due to the fact that an insufficient number of covariates are included in the record), and, conversely, two matched (identical) entities from the two data sources may give rise to different records (Fellegi and Sunter, 1969; Gu et al., 2003).

2.5.2 Propensity score matching

Besides the hard matching method, some studies apply the propensity score matching (PSM) approach in the probabilistic data linkage approach. Originally introduced by Rosenbaum and Rubin (1983), use of propensity scores has increased dramatically in the past few decades. They define propensity scores are the "conditional probability of assignment to a particular treatment given a vector of observed covariates". PSM is often used in observational studies to generate suitable control groups that are similar to treated groups when a randomized experiment is not available (Rubin and Thomas, 1996). One significant feature of PSM is that it reduces the dimensionality problem involved in multivariate analysis by reducing the matching to one constructed variable—the propensity score (Kum and Masterson, 2008).

Traditional uses of propensity score is in the field of medicine, especially the assessment of the average effect of a treatment or exposure, by estimating the probability of treatment given individual covariates such that conditioning on this probability (the propensity score) ensures that the treatment is independent of covariate patterns, and in particular by achieving balance on confounders by propensity score (Joffe and Rosenbaum 1999; Cepeda et al., 2003; Zhou and Lam, 2007). The key assumption made is that, given an exposed individual and an unexposed individual with the same (or nearly the same) propensity score, treatment assignment for these two individuals is independent of all confounding factors, and so the two observations can serve as counterfactuals for the purpose of causal inference (Westreich et al., 2010).

The same scoring method but different matching techniques are proposed as an approach to record linkage between multiple datasets (Patridge, 1998; Méray et al., 2007; Hammill et al., 2009; Fraeman, K., 2010). Each dataset is comprised by shared and unshared records, and these datasets lack a common identifier. Therefore, common multiple key identifiers are used as the covariate vector which generates the single score representing the propensity of existing in one dataset relative to the other. The propensity score can then be match-merged to link records belonging to the same record as defined by the independent fields.

The propensity score matching method ensures that any differences between the two records from two datasets are not a result of differences on the matching variables, and is ideal for making casual inferences (Rosenbaum and Rubin, 1983). It is also useful in studies with small sample sizes since when there are only a few confounding variables. However, according to Joffe and Rosenbaum (1999), propensity score matching can only control for observed confounders; that is, the propensity score cannot be counted upon to balance unobserved covariates. If there is not sufficient overlap between the two groups on the matching variables, then biases such as the regression toward the mean may occur. What's more, similar with statistical hard matching, the matching results may not be representative of the general population, since this method merely select conditional on observed common variables. Another technological issue is the calculation of propensity score by logistic regression: since there is restrict on the variance matrix of the regression, the common attributes as dummies, such as zipcodes, are not allowed to be included at one time in the regression, which may influence the accuracy of data linking.

2.5.3 Machine Learning

Since recent decades, machine learning, which can automatically detect patterns in data and then use the uncovered patterns to predict future data or to perform other kinds of decision making under uncertainty, has been used in various applications, including face recognition, voice recognition, disease diagnosis, spam email detection (Murphy, 2012).

Machine learning has been with us for a long time. Since artificial intelligence first achieved recognition as a discipline in the mid 1950's, machine learning has been a central research area. Since the late 1990s, various machine learning techniques have been developed to be used to estimate the conditional probabilities required by the Fellegi-Sunter (FS) theory (e.g., Mitchell, 1997; Friedman et al., 2003; Lu and Getoor, 2003; Taskar et al., 2003; and Zadrozny, B., 2004). These techniques are popular in that they have the advantages of allowing central supervision of processing, better quality control, speed, consistency, and better reproducibility of results (Winkler, 1995).

Back to 2000s, there exist some studies in statistics, science and other fields discussing the application of machine learning in linking records. Record linkage can be done entirely without the aid of a computer, but the primary reasons computers are often used for record linkages are to reduce or eliminate manual review and to make results more easily reproducible.

The major contribution of machine learning techniques to data linkage is the accuracy and quality of the linkage. Machine learning techniques can identify the potential patterns and especially some unobvious patterns[17] that even human expert cannot easily figure out from data. Such knowledge makes the decision in

[17] Common variables are the obvious pattern in linking record, while unobvious patterns refer to the attributes not being used in matching datasets.

data linkage wisely, instead of selecting randomly. In machine learning, the data linkage problem can be formulated as a binary classification problem, that is, judging whether two records from different datasets belong to the same entry or not. For instance, Sarawagi and Bhamidipaty (2002) and Winkler (2002) demonstrated how machine learning methods could be applied in data linkage situations where training data (possibly small amounts) were available. Larsen and Rubin (2001) and Winkler (2002) have shown that error rates can be estimated if combinations of labeled training data and unlabeled data are used in the estimation.

Conceptually traditional statistical and machine learning models are not that different. A number of advanced computational and machine learning methods generalize the original idea of parameter estimation in statistics. As mentioned by Galindo and Tamayo (2000), machine learning algorithms are more computational-based and data-driven, and by relying less on assumptions about the data (normality, linearity, etc.), is more robust and distribution-free. This algorithms not only fit the parameters of a particular model but also change the structure of the model itself and, in many cases, they are better at generalizing complex non-linear data relationships. On the other hand, however, machine learning algorithms may provide models that can be relatively large, idiosyncratic and difficult to interpret.

There are some studies which evaluated the behavior of machine learning algorithms to semi-parametric estimation of the propensity score (Setoguchi et al., 2008; Lee et al., 2010). These studies argue that the application of machine learning algorithms can help refining the propensity score matching procedure with advanced computational power. However, this application is limited due to the drawbacks of propensity score matching method. To my knowledge, there exist few studies applying machine learning techniques directly to data linkage, especially in real estate studies.

In summary, as with any linkage, the quality of the match is limited to the quality of the original data se well as the ability of the vector covariates to distinguish uniqueness among the two populations. An error-free "match" is not guaranteed. Although the match-merge process is designed to control matching error, the conclusions to be drawn from any match should ultimately rely on the quality of each data repository.

2.6 Summary

An extensive literature regarding the theoretical and empirical determinants of mortgage credit risk has developed over the past three decades (Quercia and Stegman, 1993). Literature on the possible triggers of the current financial crisis, including innovation in mortgage products, fast growth of securitization, the market players' wisdom and borrower behavior helps to get a better understanding of the determinants of credit risk and the crisis. However, as mentioned in Section 2.3, borrower behavior especially investor behavior in mortgage choices is crucial in triggering the current financial crisis but largely overlooked in previous literature. Considering the conclusion drawn from the studies of investor characteristics and expectation that among borrowers of mortgages, investors in the housing market are observed to possess several characteristics that may result in higher risks in default, it is rational to start the analysis of investor behavior. However, empirical evidence on the roles of homebuyers' expectation and investors' behavior is scarce. Firstly, most of the analyses are based on macro-level data, due to the difficulty in obtaining individual investor information. Secondly, in the housing market, it is quite difficult to distinguish investment and consumption purpose from the purchase behavior due to the heterogeneity of individuals. Due to the data limits and identification issues of investors from consumers in the housing market, it still remains an open question on how the default behavior of investors contributed to the crisis. The condominium loan market, given the distinct characteristics, provides a unique opportunity to identify and test the Case-Shiller hypothesis about the influences of investors (speculators) behavior and their expectations on mortgage market. The main analysis will be presented in Chapter 3.

Secondly, there is a large amount of evidence that foreclosures can have great influences on neighborhoods, from the view of house price decline in neighborhoods, rise in violent crime and thefts and thus the instability of the community, acceleration of racial transition, children performance, and emotional and physical impact on people. Recently, the impact of neighborhood foreclosure concentration on individual borrower's delinquency probability is increasingly emphasized and studied; those studies either support the information effects which can discourage neighboring defaults, or the contagion effect which induce more defaults. However, the impact of foreclosure concentration on the borrower's sensitivity to negative equity, i.e., to a certain extent the changing attitude of borrowers towards default option exercise, has not been fully discussed. Thus whether the information effect or foreclosure contagion effect dominates neighborhood foreclosure concentration impact on nearby borrowers' delinquency decision is an open question. These gaps motivate me to study how witnessing foreclosure signs in one's neighborhood influences someone's probability of and attitude towards exercising her mortgage default option to enter into default, which is Chapter 4 of this thesis.

When dealing with probabilistic data linkage issues in empirical studies, it raises my concern whether the current linkage method is the most appropriate one to be used. The conventional approach used in data linkage includes statistical hard matching which matches the records exactly based on common attributes among the datasets, and propensity score matching which matches the records based on the same or similar propensity scores. However, hard matching could not perform very well when facing a large number of covariates and thus could not account for selection bias resulting from limited observed covariates; and biases such as the regression toward the mean in propensity score matching may occur if there is not sufficient overlap between the two groups on the matching variables and thus the matching results may not be representative of the general population. Due to these empirical and technical constraints, and also due to the distinct characteristics of real estate data, there is a need for trying and comparing different approaches to see which approach is the most suitable for this study. In

Chapter 5, using two mortgage data as main data, I conduct a small analysis in discussing and comparing different approaches in data linkage.

Chapter 3 The Hidden Peril: The Role of the Condo Loan Market in the Recent Financial Crisis

3.1 Introduction

This study is the first to document the unique risk pattern and borrower behavior in the U.S. condominium loan market in the early 2000s. Using a representative dataset of privately securitized loans, we document that the number of condominium (condo) loan originations increased by 15-fold during the 2001–2007 period (Figure 3.1). Throughout this time period, condo loans accounted for 15% of all U.S. residential loan originations, rising from 9% in 2001 to 16% in 2007. More importantly, condo loans have also exhibited a fast growth in default rates over the years: its within-two-year default rate has increased by more than 30 times from 2003 to 2007 (Figure 3.2). These facts suggest that research has overlooked an important mortgage market segment in understanding the financial crisis.

Using a unique comprehensive loan-level dataset for private securitized loans originated during 2003–2007, we formally study the default behavior of the condominium mortgage market. The default behavior is modeled of the condo loans relative to single-family mortgages[18] using a logistic specification with a detailed list of loan and borrower characteristics, macroeconomic conditions as well as origination cohort, year and city fixed effects. In the pooled sample, a

[18] The single-family market in this paper refers to that of the detached single family houses.

condo loan, on average, is as likely to default within two years of origination as a single-family loan, after controlling for other loan and borrower characteristics. However, there is a sharp increase in condo loan defaults relative to single-family loan defaults over the years—with the most significant jump in condo loan default rates in 2006 and 2007. All else being equal, a condo loan originated in 2007 is 6.4% more likely to default within two years of origination than a single-family loan originated in the same year.

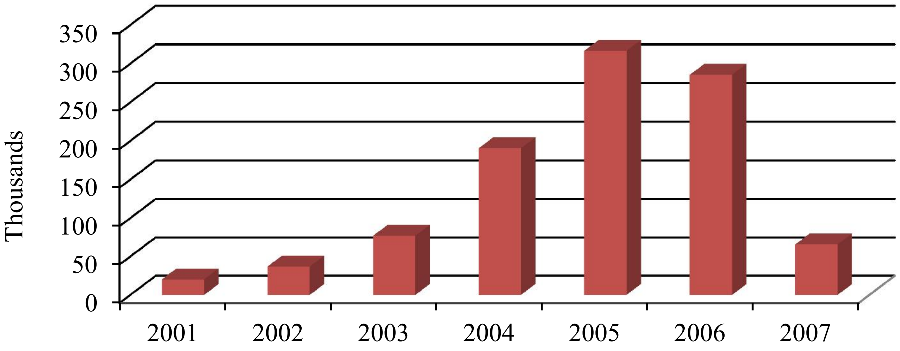


Figure 3.1 Number of condo loans originated in 2001-2007 (in thousands)

Note:

The figure shows the number of condo loan originations for all U.S. states from BBX.

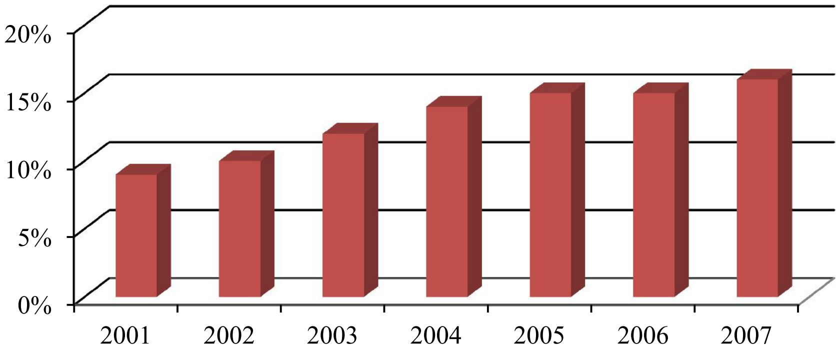


Figure 3.2 Percentage of condo loans originated in 2001–2007

Note:

The figure shows the percentage of condo loan originations for all U.S. states from BBX.

It is further shown that condo loan default rate grows at a faster rate, even compared with subprime loans—defined to be mortgages with borrower FICO score < 720—that are known to have a strong vintage effect (Demyanyk and Van Hemert, 2011). Condo loans originated before 2006, relative to single-family subprime loans originated before 2006, are much less likely to default within two years of origination. However, among loans originated in 2006, condo loans are 12% more likely to default than subprime loans in the single-family market. The pattern is even more striking if we compare subprime loans in the condo market with subprime single family loans. Given the role of the subprime market in the financial crisis, this comparison thus highlights the possibility that the much overlooked condo loan market is potentially important in understanding the crisis.

There are several competing explanations for the observed evidence on the faster default growth in the condo loan market. Condominiums and single family home markets differ in many ways. For example, compared to the detached single family houses, condominiums are typically concentrated in more urban areas. They provide different types of service flows and have distinct asset characteristics such as price growth and volatility that could lead to different default patterns over time. To address the unobserved heterogeneity issue, location fixed effect is allowed at the finer zip code level in a representative subsample and find the same results. This study also controls for direct measures of price dynamics for condominiums and single family homes using zip code level data from Zillow.com and continue to find the same pattern. Furthermore, we resort to an independent loan-level dataset from Freddie Mac. A consistent result on the Freddie Mac sample provides external validation of our previous findings. In addition, given its uniform, homogeneous nature in loan contract types (i.e., 100% of Freddie Mac loans are 30-year fully amortizing fixed-rate mortgages), the result also suggests that the observed default pattern among condo loans is unlikely driven by different loan contract terms offered to and chosen by condo borrowers.

A more interesting economic question is whether the documented default pattern among condo loans is attributable to the lender (supply side) effect or is due to the borrower (demand side) effect. Some lenders may have expertise or preference for loans in the condo market, who at the same time exhibited an increasingly lax lending standard over time. As a result, the differential default patterns between the two markets reflect the fact that disproportionately riskier borrowers are drawn to the condo loan market over time. Alternatively, condo loan borrowers could be inherently risky and thus on average default more as house prices and economic conditions deteriorate in the later years.

The results reveal that the lender channel is unlikely the driver for the observed faster growth rates of condo loan defaults relative to those of the single family loans. First, if lenders apply less stringent selection criteria over condo loan cohorts that are unobservable to econometricians, then the proportion of condo loan defaults unexplained by the observed loan and borrower characteristics should increase faster over time (compared to the single family market). This research does not find a divergence in the explanatory power of hard observable information in explaining default probabilities between the two markets. Second, Freddie Mac data is used that have information on lender identity, and the condo market's default pattern remains robust even if time varying lender fixed effects are allowed (to control for the time trend in lending standard). While there may be a common trend in credit supply that explains the default pattern for both the single family and condo loan market defaults, the supply channel is not able to explain the faster growth trajectory among condo loans. The results thus suggest the (differential) default pattern in the condo market is more associated with inherent condo borrower characteristics.

Condo borrowers are unlikely the low credit quality borrowers who default because they cannot afford to pay or refinance their mortgages as house prices start to decline. In our sample, compared to the single family market, condo borrowers have higher FICO scores, use subprime mortgages less frequently, and on average are charged a lower interest rate. A recent literature highlights the importance of homebuyer expectation and investor behavior in complementing our understanding of the financial crisis (e.g., Case et al., 2012). The price run-up in the earlier years of the decade attract more trend-chasing investors, who are more likely to make a pure economic decision of deciding to default after a significant price drop. Real estate investors likely prefer condominium units due to their smaller size (and thus smaller investment commitment), greater rental demand and less maintenance cost. Therefore, this study is likely to observe a larger presence of investors in the condo borrower population over time. In the sample, the results show a declining share of owner-occupied purchases over time, and the share of non-owner-occupied purchases is greater in the condominium market. The results are consistent with the hypothesis. Investment-purchase condo loans drive the observed condo loan market default pattern. On average, they are 30% more likely to default within two years after origination compared to other non-investment purchase condo loans, and their defaults grow at a much higher rate. Investment-purchase condo loans originated in later cohorts (e.g., 2007) are 88% more likely to default than single family loans, while the same cohort non-investment-purchase loans in the condo market are 19% less likely to default than single family loans. Default option induced by house price movements is likely key: investment purchase condo loans are 141% more likely to default when the option of default is in the money (i.e., when the house price is lower than the outstanding loan amount).

Lastly, this research explores the aggregate implication of the condo loan market defaults. It is observed that condo loan defaults have triggered more subsequent defaults in the single-family subprime market. Consistent with the notion that more of condo loan borrowers are investors who default sooner, the early default rates (i.e., within-one-year-default rates) of condo loans also grow at a faster rate than single family subprime loans. Put differently, condo loan borrowers default more promptly when house prices started to decrease in 2006 and 2007. More importantly, at the zip code level, first-year defaults among condo loans positively predict second-year defaults of the same cohort's single-family subprime loans. Early condo loan defaults also predict negatively subsequent single family house price growth at the same zip code. Granger causality tests verify the temporal lead-lag relationship: condo defaults precede the subprime mortgage defaults as well as the house price growth in the single family market. This provides new insight on the triggering event of the housing crisis. To better identify the channel, this work uses the exogenous variation in state foreclosure laws and find the effect of condo loan defaults, on both the subsequent house price growth and the subsequent single family subprime defaults, to concentrate in the non-judicial states in which foreclosure process is more efficient. These results are consistent with the idea that condo defaults prompt more defaults of single family subprime mortgages at the same location, through the channel that foreclosures on the defaulted properties depress neighboring house prices.

This study contributes to the literature by first documenting a strong, robust and economically important default pattern in the much ignored condominium loan market. Specifically, the loan origination growth rate and default pattern in the condo market are comparable to the subprime mortgage market (Demyanyk and Van Hemert, 2011). The findings in this chapter also add to the understanding of the economic channels that explain the financial crisis. A large strand of the literature has focused on the subprime mortgages[19] and other supply side factors such as the role of securitization.[20] On the other hand, recent work (Haughwout, et

[19] These products were designed to help borrowers in markets expecting significant price appreciation. However, they were often marketed to borrowers with relatively poor credit histories as well. As a result, these mortgages are often referred to as subprime mortgages, because they did not meet the underwriting criteria set by the government-sponsored enterprises, e.g. Fannie Mae and Freddie Mac (See Agarwal, Ambrose, Chomsisengphet, and Sanders 2012).

[20] For a discussion, see Agarwal et al. (2011); Agarwal, Chang, and Yavas (2012); Agarwal and Evanoff, 2013; An, Deng, and Gabriel (2011); Jiang, Nelson, and Vytlacil (2012); Keys, Mukherjee, Seru, and Vig (2010a, 2010b); Mian and Sufi (2009); Mayer, Pence, and Sherlund (2009); Piskorski, Seru, and Vig (2010); and An, Deng, Rosenblatt, and Yao (2012).

al., 2011; Case et al., 2012) suggests that a less studied but potentially fundamental factor may have triggered the crisis—homebuyer and especially investor expectations. Based on a survey sample in four U.S. cities, Case et al. (2012) report that home price expectations, which reached abnormal levels relative to the mortgage rate at the peak of the boom and have declined sharply since 2007, were highly correlated with the price movements of the housing market. Cheng et al. (2013) find that mid-level managers in securitized finance business continue to speculate on house prices in their own home purchases during the boom period. Using transaction-level data, recent studies find supportive evidence of housing speculators chasing short-term trends (Bayer et al., 2011), leading to price overreaction (Chinco and Mayer, 2014; Fu and Qian, 2013). Haughwout et al. (2011) use unique credit report data to show the important role speculative investors play in contributing to the rise and fall of the U.S housing market in the recent crisis. Specifically, when prices turned downwards, these investors defaulted in large numbers, contributing to the intensity of the housing cycle's downward leg. Amromin et al. (2011) identify another demand-side factor: high credit worth households chose complex mortgage products leading up to the crisis and these households were more likely to default. Other work documents borrowers with unconventional mortgages, or who misrepresented their financial network are risky borrowers and default more (e.g., Garmaise, 2013a, 2013b). The findings in this chapter complement the demand-side view by providing evidence that investor behaviour, as manifested in the condo market's loan default pattern in our context, play an important role in explaining mortgage defaults in the crisis. In addition, this study relates to the prior literature on the real effect of the housing crisis (e.g., Campbell, et al., 2011; Mian, et al., 2012) and shows that condo loan market defaults have aggregate implications on the house prices and default patterns in other segments of the housing market.

The rest of this chapter proceeds as follows. The next section reviews the literature in mortgage market and financial crisis and highlights the contributions of this chapter to the previous studies on credit risks and financial crisis. The data used for this study is described in Section 3.2, and the hypotheses and empirical methodology are shown in Section 3.3. In Section 3.4, the empirical results that document the condo market's default pattern are presented. Section 3.5 performs analysis to differentiate competing economic explanations. Next the aggregate implication of condo loan defaults is presented by studying its spillover effects in Section 3.6. Finally, we make concluding remarks in Section 3.7.

3.2 Data Description

3.2.1 Data sources

The first and primary source for the study is the loan-level data furnished by BlackBox Logic (BBX). The loan-level data from BBX cover loans originated in 2003–2007 (we leave out loans originated earlier due to better data coverage in the later sample period). BBX aggregates data from mortgage servicing companies that participate in their servicing agreement. The most recent BBX data cover about 22 million mortgages throughout the United States, making it a comprehensive source for both the prime and subprime mortgages.[21] For example, based on a comparison with HMDA data that include a near complete universe of U.S. mortgage applications and originations, The BBX data is estimated to cover about 70% of the prime market during the period. A representative sample of the subprime mortgage market allows us to compare the default behavior of the condominium loans with that of the subprime mortgages that plays a key role in triggering the financial crisis.

In addition to monthly data on loan performance, BBX contains information on borrower and loan characteristics at origination, including the borrower's FICO credit score, the loan amount and interest rate, whether the loan is a fixed- or adjustable rate mortgage, LTV, and whether the loan was intended for home purchase or refinancing, among other characteristics. The outcome variable that we focus on is whether the loan becomes 60 days or more past due within the 24 months following origination. The BBX loan-level data is also merged with macro variables, including the slope of the yield curve and the credit spread from Federal Reserve Bank of St. Louis, the state-level unemployment rate from Bureau of Labor Statistics, and the MSA-level quarterly purchase-only housing index from the Office of Federal Housing Enterprise Oversight (OFHEO), which was succeeded by the Federal Housing Finance Agency (FHFA).[22]

[21] BBX define subprime mortgages as those with borrower FICO score < 720.

[22] Established in 2008, FHFA is a successor agency that resulted from the statutory merger of the Federal Housing Finance Board (FHFB), the Office of Federal Housing Enterprise Oversight

Despite its market coverage as well as richness in many loan and borrower characteristics, the BBX dataset has limitations. For example, it does not contain lender information. In the later analysis where we examine sources of the default pattern in the condominium market, a second dataset is used – loan-level performance data from Freddie Mac. This dataset, recently made publicly available, includes loan-level origination and monthly loan performance data on approximately 15.7 million fully amortizing 30-year fixed-rate mortgages that Freddie Mac acquired. While the Freddie Mac sample does not cover the subprime mortgage market, it serves as a great supplementary dataset for this analysis for the following reasons. First, Freddie Mac loan-level data contain lender identity information, which allows us to differentiate between the borrower channel and the lender-related effect. Second, loans in the Freddie Mac sample are fixed-rate 30-year agency mortgages with full documentation. This stands in contrast to the BBX sample where condo loans are much more likely to have exotic contract terms and have little or no documentation. Analysis based on this more homogeneous sample of loans thus helps establish robustness of our results and distinguish from alternative interpretations. In addition, the Freddie Mac data allow us to explore in depth the specific channel of the observed default pattern in the condo market. To ensure consistency on the condo vs. single family loans comparison across the two datasets, we note that both datasets have a comprehensive coverage of loans in their focus markets and the fractions of condominium loans are comparable (14% in BBX vs. 11% in Freddie Mac).

(OFHEO), and the U.S. Department of Housing and Urban Development's government-sponsored enterprise mission team.

3.2.2 Descriptive statistics of condo versus single-family loans

This analysis keeps loans in the BBX dataset that were originated between 2003 and 2007 in the single-family and condominium markets. It also restricts to purchase loans with original loan balances smaller than $10 million. Since condos likely concentrate in larger urban areas, we restrict our sample to the top 2000 cities, which cover 98% of the entire condo market in the BBX dataset. Each zip code is further required to have more than 5 loans originated per year to be included in the sample.[23] The final sample contains 5,000,241 observations, with 909,564 condo loan observations (18.2%) and 4,090,677 single family loans observations (81.8%). [24]

Table 3.1, Panel A shows summary statistics of the major variables in the pooled sample for the 2003–2007 period. Among all the loans, 40% are fixed-rate mortgages and 26% are subprime mortgages. Borrowers have an average FICO credit score of 683, and take out up to 73% of the property value (LTV). Overall, the probability of default within two years of loan origination is 6% on average.

Loan and borrower characteristics in the condo and single-family home mortgage market differ. On the one hand, condo loans appear safer along many dimensions. Condo borrowers have higher FICO credit scores than single-family borrowers (by 20 points). The number of subprime loans in the condo market is

[23] Using this data screening to remove zipcodes with less than 5 loans originated per year, only 0.8% of observations have been eliminated.

[24] We follow the same rules to construct the sample from our supplementary data source Freddie Mac and for brevity we leave the summary statistics for the Freddie Mac sample in the Appendix.

one-third smaller than in the single-family loan market. The average condo borrower's interest rate is significantly lower than that of single-family borrowers.

On the other hand, condo loans typically involve less conventional contract terms. In the condo loan market, this study observes much fewer fixed-rate mortgages and considerably more option ARMs, interest-only loans, and low or no documentation loans than in the single-family market. In addition, fewer condo borrowers purchase for owner-occupancy, and they tend to buy in more expensive areas (i.e., those with a higher FHFA/OFHEO House Price Index, or HPI).

Table 3.1 Summary Statistics of BlackBox Full Sample

Panel A: Summary statistics for BlackBox (BBX): from 2003 to 2007				
	Total	**Condo**	**Single-Family (SF)**	**Diff. (Condo-SF)**
D_default within 2 yrs	6%	4%	6%	-2%***
FICO score	683	699	679	20***
Original LTV	73%	73%	73%	0***
Original loan balance (*1000)	230	220	232	-12***
Current interest rate	7.45	7.11	7.53	-0.42***
Margin	2.27	2.21	2.28	-0.07***
D_Subprime	26%	19%	27%	-8%***
D_FRM	40%	35%	41%	-6%***
D_Second lien	19%	18%	20%	-2%***
D_Option ARM	5%	8%	4%	4%***
D_Interest only loan	22%	29%	21%	8%***
D_Heloc	2%	2%	2%	0%***
D_Low/No doc	35%	41%	34%	7%***
D_Owner occupied	73%	69%	74%	-5%***
Log_HPI	5.33	5.36	5.32	0.04***
Log_duration	6.86	6.94	6.84	0.10***
Sample Size (in thousands)	5,000	909	4,091	

Panel B: Summary statistics for BlackBox (BBX) by loan origination year (2003–2007)

	Condo	Single-Family	Diff.	Condo	Single-Family	Diff.	Condo	Single-Family	Diff.
Original Year		**2003**			**2004**			**2005**	
D_default within 2 yrs	0%	1%	-1%***	1%	1%	0%***	2%	4%	-2%***
FICO score	697	683	14***	701	680	21***	700	679	21***
Original LTV	73%	73%	0%	74%	75%	-1%***	73%	73%	0%***
Original loan balance （*1000）	212	234	-22***	211	221	-10***	214	224	-10***
Current interest rate	6.93	7.26	-0.33***	6.69	7.22	-0.53***	6.93	7.47	-0.54***
Margin	1.58	1.72	-0.14***	2.21	2.29	-0.08***	2.45	2.52	-0.07***
D_Subprime	16%	23%	-7%***	19%	28%	-9%***	20%	29%	-9%***
D_FRM	38%	46%	-8%***	27%	33%	-6%***	32%	38%	-6%***
D_Second lien	14%	14%	0%	13%	14%	-1%***	18%	21%	-3%***
D_Option ARM	1%	0%	1%***	5%	3%	2%***	9%	5%	4%***
D_Interest only loan	11%	5%	6%***	25%	17%	8%***	34%	25%	9%***
D_Heloc	0%	0%	0%	1%	1%	0%***	3%	3%	0%***
D_Low/No doc	24%	24%	0%***	29%	25%	4%***	42%	35%	7%***
D_Owner occupied	78%	82%	-4%***	75%	78%	-3%***	68%	74%	-6%***
Log_HPI	5.41	5.34	0.07***	5.43	5.38	0.05***	5.37	5.33	0.04***
Log_Duration	6.85	6.85	0	6.86	6.77	0.09***	6.96	6.86	0.1***
Sample Size (in thousands）	75	441		185	901		317	1,758	

Panel B: Summary statistics for BlackBox (BBX) by loan origination year (2003–2007) (Continued)

	Condo	Single-Family	Diff.	Condo	Single-Family	Diff.
Original Year		**2006**			**2007**	
D_default within 2 yrs	9%	12%	-3%***	10%	13%	-3%***
FICO score	694	674	20***	710	689	21***
Original LTV	70%	71%	-1%***	76%	79%	-3%***
Original loan balance （*1000）	212	228	-16***	324	325	-1
Current interest rate	7.58	7.94	-0.36***	7.32	7.46	-0.14***
Margin	2.24	2.38	-0.14***	1.58	1.55	0.03
D_Subprime	20%	29%	-9%***	9%	18%	-9%***
D_FRM	41%	45%	-3%***	47%	51%	-4%***
D_Second lien	24%	25%	-1%***	16%	16%	0%
D_Option ARM	9%	5%	4%***	10%	5%	5%***
D_Interest only loan	31%	24%	7%***	34%	26%	8%***
D_Heloc	2%	2%	0%***	0%	0%	0%
D_Low/No doc	48%	40%	8%***	53%	43%	10%***
D_Owner occupied	65%	69%	-4%***	60%	61%	-1%**
Log_HPI	5.31	5.27	0.04***	5.3	5.25	0.05***
Log_Duration	6.98	6.86	0.12***	6.96	6.9	0.06***
Sample Size (in thousands）	278	1,159		63	265	

Panel C: Comparison of Single Family Subprime Loan Market and Condo Loan Market (Full and Subprime)

	SF Subprime	Condo	Diff. (Condo-SF Subprime)	Subprime Condo	Diff. (Subprime Condo-Subprime SF)
D_default within 2 yrs	7%	4%	-3%***	6%	-1%***
FICO score	578	699	122***	579	1***
Original LTV	78%	73%	-6%***	77%	-1%***
Original loan balance (in thousands)	165	220	68***	161	-4***
Current interest rate	8.22	7.11	-1.19***	8.19	-0.03***
Margin	3.74	2.21	-1.59***	3.86	0.12***
D_FRM	29%	35%	6%***	28%	-1%***
D_Second lien	11%	18%	7%***	10%	-1%***
D_Option ARM	0%	8%	8%***	0%	0%***
D_Interest only loan	9%	29%	21%***	12%	3%***
D_Heloc	1%	2%	1%***	3%	2%***
D_Low/No doc	14%	41%	27%***	17%	3%***
D_Owner occupied	80%	69%	-11%***	80%	0%**
Log_HPI	5.29	5.36	0.09***	5.36	0.07***
Log_Duration	6.80	6.94	0.06***	6.85	0.05***
Sample Size (in thousands)	1,123	909		170	

Note:

This table presents the summary statistics of the BlackBox Logic (BBX) dataset. This dataset includes only single-family and condominium (condo) loans originated during the period 2003–2007. Panel A reports the results from aggregate-level summary statistics of the loans and compares the average values of the variables by full sample, single-family loans, and condo loans, respectively. Panel B shows the full sample summary statistics results by origination year. Panel C shows the comparison of the single family subprime loan market and the condo loan market (full and subprime). D_default within 2 yrs is equal to one for defaulting within two years of the loan origination date. Current interest rate refers to the coupon rate charged to the borrower for the most recent remittance period. Original loan balance is defined as the amount of principal on the closing date of the mortgage. FICO score refers to the FICO (formerly the Fair Isaac Corporation) borrower credit score at the time of the loan closing. Original LTV means the ratio of the original loan amount to the property value at loan origination. D_FRM is equal to one for fixed-rate mortgages. D_Owner occupied takes one if the property is owner occupied. D_Second lien is equivalent to one for a second lien loan that is subservient to the main or first mortgage on a piece of real estate. D_Option ARM is equal to one if it is an adjustable rate mortgage with added flexibility of making one of several possible payments on your mortgage every month. D_Interest only loan is one if it is a loan in which, for a set term, the borrower pays only the interest on the principal balance with the principal balance unchanged. D_Heloc is equivalent to one if it is a loan in which the lender agrees to lend a maximum amount within an agreed term, where the collateral is the borrower's equity in his/her house (HELOC is short for home equity line of credit). D_Low/No doc takes one if the borrower is required to provide low or no documentation. D_Subprime equals one if it is a subprime loan (i.e., loans with FICO score lower than 720). Margin is the difference between the interest rate charged to the borrower and the applicable ARM index, as measured in number of percentage points. Log_HPI is log of the quarterly FHFA/OFHEO House Price Index. Log_Duration is the log of the elapsed time from origination to the end of the sample period or to the first classification as being prepaid or delinquent at least 60 days. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

To further examine the differences between the two markets among different origination cohorts, the pooled sample averages is decomposed by each origination year cohort for both the condo and single-family loans in the period 2003–2007 (Table 3.1, Panel B). The number of both single-family and condo loans peaked in 2005 and then sharply declined in 2007. The number of risky loan contracts such as option ARMs, interest-only rate mortgages, and low or no documentation mortgages have risen over the years, and the increase is faster among condo loans than among single-family loans. On the other hand, the difference in the fraction of subprime mortgages between the condo and single family loans remains steady during the four year period. The FICO score difference even increases: the average condo borrower has an even higher FICO score than the average single family borrower in the later cohorts of the sample period. In addition, the gap in non-owner-occupied purchases between the condo and single family loans shrinks: owner-occupancy decreases in both markets but the decrease is faster in the single family loan market. To the extent that loan contract terms, borrower credit worthiness, and owner occupancy status potentially capture different aspects of the risk associated with a mortgage, it remains ambiguous whether the loans in the condo market become more or less risky in the later origination cohorts relative to the single family loans.

Next the loan and borrower characteristics in the condo loan market are compared to the subprime segment of the single family market (Table 3.1, Panel C). Since subprime mortgages are riskier loans than prime mortgages, the comparison between condo loans and subprime loans is more informative about the characteristics and risk profiles of condo loans. Most of the differences in the loan and borrower characteristics observed in the full sample (Table 3.1, Panel A) remain to hold. Condo borrowers are much more creditworthy by the conventional measures. However, even compared to the subprime mortgages in the single family market, condo loans are much more likely to have non-conventional contract terms or have low or no documentation. Similarly, the condo purchases are less likely to be owner-occupied. This essay also studies the borrower and loan characteristics in the subprime segment of the condo loan market in comparison to those in the subprime segment of the single family market. The subprime market is relatively homogeneous across the condo and single-family markets: although we still observe a robust pattern of riskier loan terms in the condo subprime market, the differences in borrower and loan characteristics between these two markets are much smaller than between all condo loans and single-family subprime loans. This suggests that it is not simply subprime mortgages that explain the distinct characteristics of the condo loan market. On the contrary, condo loans in the prime market are particularly associated with riskier contract terms, low or no documentation, and are more likely to have risky borrowers (e.g., investors).

3.3 Hypotheses and Methodology

3.3.1 Hypothesis Development

Hypothesis 1

The first null hypothesis is that the condo and single-family markets are similar. Hence, the default rates (as well as default growth patterns) for condo and single-family loans should exhibit similar patterns. This hypothesis is tested to determine if condo loans have higher default rates by focusing on comparable single-family and condo markets.

Hypothesis 2

The Driving factor of the unique default pattern in the condo loan market is neither due to the unobserved factors in condominium market compared to the single family market, not due to the lender preference and/or expertise with loans in the condo market and the single family market. The unique default pattern in condo loan market is mainly because condo loan borrowers could be inherently riskier, compared with single family loan borrowers. Given that real estate investors are more present in the condominium market, the observed default pattern in the condo loan market thus may be associated with the investor behavior.

Hypothesis 3

The impact of condo loan defaults resulting from risky borrowers has negative spillover effects on the neighboring single-family market. If this hypothesis holds, not only should condo loans default earlier compared with single-family loans originated in the same cohort, but also the earlier condo loan defaults prompt more defaults in the single-family sector in the same area afterwards. Therefore, the question can be asked do early condo defaults predict the single family subprime market's subsequent default rate.

3.3.2 Methodology

The main empirical specification in this study is a logistic model of the default decision of loans originated between 2003 and 2007.[25] A loan to be in default is defined if it becomes delinquent by at least 60 days[26] within two years of origination. The main independent variable, Condo$_{is}$, is a binary variable that is set to one if the loan is a condo mortgage. Other explanatory variables include both loan-level and macro-level variables. City, and year fixed effects are included to control for unobservable factors at the city level and at the year level. Loan *i* enters the study in month t_{is}, which is the first occurrence of that loan. The same loan exits the study in month T_{is}, which is the earliest occurrence of one of the "exit" events (default or prepay or the end of the sample period). Finally, all the standard errors reported in main default analysis, unless otherwise stated, are clustered at the city level, in addition to being robust to heteroskedasticity.

[25] For robustness, we replicate our analysis using a linear probability model and find consistent results as well.

[26] More specifically, we define default as a loan that is delinquent by at least 60 days, or that is in foreclosure, is in bankruptcy, is REO (real estate owned), or is in the liquidation stage.

Loan-level controls are motivated by the literature. They include indicators for FICO credit scores, indicators for fixed-rate and interest-only loans, indicators for low- and no-documentation (low/no doc) loans, an indicator for owner-occupancy status, an indicator for subprime mortgages, and an indicator for home equity lines of credit (HELOC). Following the literature, this study also includes an indicator variable for LTV at origination of 80% as a proxy for the existence of a second lien on the property. Continuous loan-level variables include (log of) the loan amount, the first interest rate observed, the time elapsed from origination to the end of the sample period or to the first classification as being prepaid or delinquent at least 60 days, and LTV at the time of origination. This analysis also includes the current level of the residential home price index, the state-level unemployment rate, the slope of the yield curve, and the credit spread as control variables.

3.4 Empirical Analysis on Condo Default

3.4.1 Unconditional result of the default behavior of the condo market

The default rates within two years of origination in both markets increased over the years in our sample (Figure 3.3.1). More importantly, the increase in the condo loan default rate is much faster. Among the 2003 cohort loans, the default rate in the single-family market is more than double that of the condo market.

However, among loans originated in 2007, the two-year condo default rate is 10.1%, which is comparable with 12.6% in the single-family market.

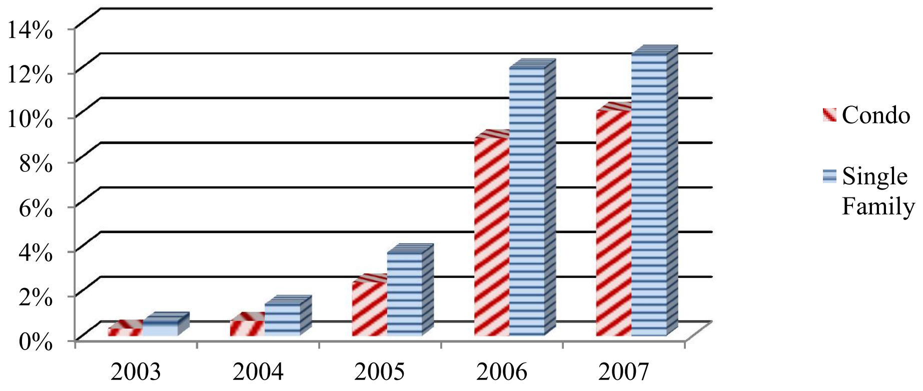


Figure 3.3 Frequency Distribution of Defaults within Two Years of Origination: Condo vs. Single-Family

Fig. 3.3.1 Frequency distribution of defaults within two years: Full sample

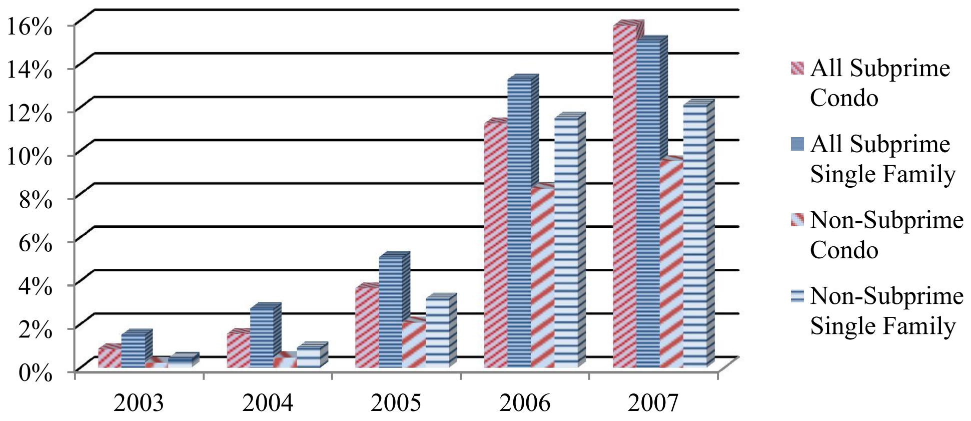


Fig. 3.3.2 Frequency distributions of defaults within two years: Subprime and non-subprime

Note:

This figure shows the frequency distribution of loan defaults within two years of origination (in percentages). All the loans are originated during the period 2003–2007 and are separated by property type: condo and single-family. Fig. 3.3.1 shows the frequency distribution of within-two-year default rates for the full sample; Fig. 3.3.2 presents the distribution by comparing subprime and non-subprime loans. The Y-axis indicates the percentage of default probability within two years of origination, and the X-axis indicates the origination year of the loan.

Figure 3.3.2 shows a decomposition of the default patterns in the condo and single-family markets, by subprime and non-subprime status. Within the subprime and non-subprime sub-markets, condo loan defaults start at a much lower rate than single-family loan defaults, but grow more quickly over the sample period. Specifically, the rate of condo subprime loan defaults exceeds that of the single-family market by 0.7 percentage points among loans originated in 2007. These results, in combination with our previous findings, imply that condo loans have distinct features that make them riskier and more vulnerable to default, especially during times of market distress.

3.4.2 Regression analysis of condo loan default behavior

Option-based theoretical and empirical models for mortgage default analysis have been well developed during the past two decades (e.g., Kau et al., 1992; Kau and Keenan, 1999; Deng et al., 1996, 2000), and they have increased in realism and sophistication in the past decade (e.g., Ambrose et al., 2001; Deng and Gabriel, 2006). Clapp et al. (2006) provide a comprehensive review of these modeling frameworks. Following Clapp et al. (2006), we perform logistic regressions to formally study the default behavior of the condo market relative to the single-family market. Because condo loans differ substantially from single-family loans in their loan and borrower characteristics in the BBX dataset, observables on loan and borrower characteristics are included as controls in the logistic analysis. Macroeconomic variables are also included as well as origination cohort, year, and city fixed effects in the regression. Table 3.2 reports odds ratios in the full sample analysis: an odds ratio greater (smaller) than one indicates a positive (negative) effect. Consistent with the literature, FICO scores, LTV, FRM loan type, and owner-occupancy status are strong predictors of default. Second lien loans and low/no doc loans are risky, as they are associated with higher default rates within two years of origination.

Although condo loans have a lower average default rate in the summary statistics (Table 3.1, Panel A), the logistic analysis of Table 3.2 shows that after controlling for loan and borrower characteristics, condo loans do not differ much from single-family loans in their two-year default probability. The Condo dummy coefficient is economically and statistically indistinguishable from 1 (i.e., the odds of observing a condo loan default are as high as observing a single family loan default). Furthermore, separating default behavior by origination year reveals a significant time trend in the condo market defaults that is consistent with the time-series pattern shown in Panel B of Table 3.1 and Figure 3.2. Coefficient on the interaction term between the condo dummy and the origination year *t* captures the difference in odds ratio between condo loans' and single family loans' default rates in origination year *t,* relative to the odds ratio difference between the two submarket's loan defaults in the origination year 2003 (i.e., the absorbed origination year). Therefore, those coefficients in Table 3.2, which are greater than one, suggest that there is a sharper increase in condo loan defaults relative to single-family loan defaults over the years—with the most significant jump in condo loan default rates in 2006 and 2007. As a result, condo loans default more than single family loans in later cohorts. While condo loans originated in 2003 are

44% less likely to default, condo loans originated in 2007 are 6.4%[27] more likely to default within two years of origination than single-family loans in the same cohort.

Table 3.2 Logistic Analysis of Borrower within-two-year Default: Condo vs. Single-Family

	(1) D_default within 2yrs	(2) D_default within 2yrs
D_Condo	0.992	0.562***
	(-0.26)	(-5.97)
D_Condo * D_OrigYear2004		1.097
		(0.84)
D_Condo * D_OrigYear2005		1.536***
		(4.30)
D_Condo * D_OrigYear2006		1.920***
		(6.18)
D_Condo * D_OrigYear2007		1.894***
		(5.06)
D_Owner occupied	0.595***	0.595***
	(-30.43)	(-30.43)
D_Second lien	6.834***	6.822***
	(44.02)	(44.07)
D_FRM	0.706***	0.705***
	(-23.69)	(-23.80)
D_Option ARM	0.441***	0.440***
	(-13.90)	(-13.86)
D_Interest only loan	0.965*	0.965*
	(-1.78)	(-1.81)
D_Heloc	0.620***	0.618***
	(-17.54)	(-17.70)
D_Low/No Doc	1.023	1.021
	(1.52)	(1.40)
D_Subprime	0.739***	0.738***
	(-16.95)	(-16.94)
Original LTV	1.002***	1.002***
	(18.05)	(18.04)
Log_FICO score	0.008***	0.008***
	(-45.87)	(-45.95)
Log_Original loan balance	1.383***	1.382***
	(13.90)	(13.94)
Log_HPI	0.290***	0.289***
	(-7.24)	(-7.27)
Log_Duration	0.149***	0.149***
	(-63.63)	(-63.42)
Unemployment rate	0.968	0.967
	(-1.56)	(-1.63)
Yield slope	1.153***	1.154***
	(6.38)	(6.46)
Credit spread	1.635***	1.636***
	(49.42)	(49.47)

[27] The coefficient on condo dummy is multiplied with the coefficient on condo x origination year 2007 interactive term (0.562 x 1.894 = 1.064) to compute the odds ratio of a 2007-originated condo loan default relative to a 2007-originated single family loan default.

Table 3.2 Logistic Analysis of Borrower within-two-year Default: Condo vs. Single-Family (Continued)

	(1) **D_default within 2yrs**	(2) **D_default within 2yrs**
Current interest rate	1.049***	1.048***
	(14.75)	(14.61)
Fixed effects	City, current year and origination year	
Observations	2,291,374	2,291,374
Pseudo R-squared	0.418	0.418

Note:

This table presents the result of logistic regression analysis for the refined 2000-city BBX sample. This dataset includes only single-family and condominium (condo) purchase loans (<10million USD in origination amount) from all states originated during the period 2003–2007. The dependent variable D_default within 2 yrs takes the value of one for defaulting within two years of the loan origination date. The definitions of the independent variables are shown in Table 3.1. Standard errors are clustered at city level. City fixed effects, current year fixed effects and origination year fixed effects are included in the regression but not reported. Odds ratios are reported and robust z-statistics are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

A subsample analysis is performed of all condo loans and all single-family subprime loans. The result in Column (1) of Table 3.3 show that, on average, condo loans are slightly less likely to default compared to single family subprime mortgages. In earlier vintages, single-family subprime loans are consistently more likely to default than condo loans of the same vintage. To the extent that condo loans are compared to a riskier segment of the mortgage market, the test is constructed against finding a significant result. However, condo loan defaults grow at a faster rate and over time condo loans begin to default more than single-family subprime loans (Column (2) of Table 3.3). Condo loans originated in 2006 are 12% more likely to default within two years of origination than single-family subprime loans originated in the same year, after controlling for all the observed loan and borrower characteristics. Given the role of the subprime mortgages in the financial crisis, this comparison thus highlights the possibility that the much overlooked condo loan market is potentially important in understanding the crisis.

The pattern becomes more apparent when we compare condo and single family loans within the subprime market. In this sample, condo subprime loans are riskier than single-family subprime loans—a condo subprime loan is 13% more likely to default than a single-family subprime loan (Column (3) of Table 3.3). The higher default probability among condo subprime loans is driven by later vintages. Condo subprime loans originated in 2006 and 2007 are 29% and 81% more likely to default than single-family subprime loans originated in the same years.[28]

[28] For completeness, we also perform a direct comparison, using the same specification as in Table III, of two year default rates between (a) prime condo loans with prime single family loans; and between (b) prime condo loans with subprime single family loans. We find that over time prime condo loans exhibit a higher growth rate in defaults than prime single family loans. On the other hand, condo prime loans are less risky than single family subprime loans (for all origination cohorts), which is consistent with the findings in the literature on the default risk difference between prime and subprime loans.

Table 3.3 Logistic Analysis of within-two-year Default: Condo Loans vs. Subprime Loans

	(1)	(2)	(3)	(4)
	D_default within 2yrs		**D_default within 2yrs**	
	All Condo vs SF Subprime Loans		**Condo Subprime vs. SF Subprime Loans**	
D_Condo	0.895***	0.301***	1.125***	0.758*
	(-3.20)	(-10.90)	(3.90)	(-1.81)
D_Condo x		1.317**		1.002*
D_OrigYear2004		(2.44)		(0.01)
D_Condo x		2.445***		1.151
D_OrigYear2005		(7.95)		(0.89)
D_Condo x		3.724***		1.700***
D_OrigYear2006		(10.95)		(3.35)
D_Condo x		3.016***		2.384***
D_OrigYear2007		(7.76)		(4.56)
Controls	Yes	Yes	Yes	yes
Fixed effects	City, current year and origination year		City, current year and origination year	
Observations	742,517	742,517	402,533	402,533
Pseudo R-squared	0.373	0.375	0.333	0.334

Note:

This table presents the result of logistic regression analysis that includes all condo loans and subprime loans originated during the period 2003–07 in the BBX sample. Columns (1) and (2) present the logistic regression results of all condo loans and single-family subprime loans, and columns (3) and (4) show results of condo subprime loans and single-family subprime loans. The dependent variable *D_default within 2* yrs takes the value of one for defaulting within two years of the loan origination date. We include the same set of control variables as in Table 3.1. City fixed effects, current year fixed effects and origination year fixed effects are included in the regression but not reported. Standard errors are clustered at city level. Odds ratios are reported and robust z-statistics are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

3.5 What Drives the Default Pattern in the Condo Loan Market?

Condo markets differ from the single family in many aspects. As a result, there are several competing explanations for the observed evidence in Table 3.2 and 3.3 that the condo loan default rate grows at a faster speed and its level eventually exceeds that in the single family market, including the riskier subprime market segment. Compared to the single family market, condominiums are typically concentrated in more urban areas. They provide different types of service flows and have distinct asset characteristics such as price growth and volatility that could lead to different default patterns over time. Alternatively, lenders have different preference and/or expertise with loans in the condo market and the single family market, and the differential default patterns between the two markets reflect the lender (or supply-side) effect. Lastly, condo loan borrowers could be inherently riskier. Several approaches are used to distinguish these explanations.

3.5.1 Unobserved characteristics of the condo market

First, a subsample analysis is performed to better control for the location fixed effects. In the previous analysis, any time-invariant characteristics at the city level are removed. However, finer geographical boundaries are needed to address the concern that condo loans are simply located in different, potentially riskier areas. To do so, zip code fixed effects are included in an analysis based on the subsample of the top 50 cities in our sample (based on the number of total loans in our entire period). The choice of the subsample analysis (instead of full sample regression) is motivated by the following considerations. There are 6,914 zip codes in our sample, which imposes significant computational challenge in our logistic regression. In addition, the top 50 city subsample constitutes 34% of the entire condo loans, which is fairly representative of the full sample. Column (1) of Table 3.4 presents the results using finer zip code fixed effects based on the subsample analysis. The result exhibits the same pattern as before. Condo loan defaults are increasing over the origination cohort years, and in particular, condo loans originated in later cohorts (e.g., 2006) default more than their single family counterparts.

Condominiums arguably provide different types of service flows to owners, compared to detached single family houses. As a result, the two types of housing markets will have distinct asset characteristics (e.g., growth rate and volatility). The option pricing theory implies that the observed differences in mortgage defaults could result from differences in the underlying asset volatility. Therefore, if condo properties have a different return generating processes from single family properties, then this would lead to the observed differences in default risk between condo and single family mortgages. The analysis in Table 3.2 and 3.3 includes MSA-level HPI index, which is primarily based on transactions in the single family market, and is thus not able to address this issue. This possibility is examined by including specific asset characteristics for the condominium and single family market respectively. Specifically, this analysis is conducted in a subsample where there is detailed information about price dynamics in the condominium market using Zillow house price data which contain monthly transaction price at the zip code level for the condo and single family markets separately. Consistent with the analysis in Column (1) of Table 3.4, the top 50 cities subsample after merging Zillow with BBX is the focus, which covers 33% of the entire condo loans in the sample.

The analysis includes the zip code-level average transaction price and the growth rate of the average transaction price for both the condo and the single family market in the regression to control for the asset market dynamics in these two markets (Column (2), Table 3.4). The zip code fixed effects are also used to control for any unobserved location effects at the zip code level. The results continue to hold. In Column (3) of Table 3.4, location-specific time trend is further included in the analysis to better control for the dynamics of local markets (e.g., local trend in supply of condominiums relative to the single family houses).[29] The same results still remain: condo loans default faster and their later cohorts default more than single family loans in the same cohort, after controlling for asset dynamics in the two markets and a location-specific time trend.[30]

[29] In the specifications that we include location-specific time fixed effects, we modify our standard error clustering at the location time level. This is to allow correlation in defaults within each location at a given year.

[30] The weaker effect after we control for asset market characteristics suggests that the difference in asset attributes may account for part of the difference in default patterns. It may also be due to lack of power as we restrict to a more limited sample in this analysis.

Table 3.4 Controlling for Location and Asset Dynamics of the Condo Market

	(1)	(2)	(3)
	D_default within 2yrs	**D_default within 2yrs**	**D_default within 2yrs**
	BBX subsample (top 50 cities)	**Zillow-BBX merged sample (top 50 cities)**	
D_Condo	0.900	0.947	0.945
	(-0.73)	(-0.31)	(-0.38)
D_Condo * D_OrigYear2004	0.858	0.866	0.843
	(-0.83)	(-0.71)	(-0.80)
D_Condo * D_OrigYear2005	1.214	1.103	1.074
	(1.24)	(0.50)	(0.40)
D_Condo * D_OrigYear2006	1.558**	1.445*	1.358*
	(2.55)	(1.76)	(1.85)
D_Condo * D_OrigYear2007	1.332*	1.328	1.257
	(1.34)	(1.09)	(0.93)
Log_Condo price level		1.971**	1.090
		(2.32)	(1.34)
Log_Single family price level		1.276	0.472***
		(0.59)	(-9.47)
Condo price growth		0.078***	0.227**
		(-4.30)	(-2.28)
Single family price growth		0.000***	0.000***
		(-7.87)	(-5.30)
Controls	Yes	Yes	yes
	zip codes, current year	zip codes, current year	
Fixed effects			City-year and origination year
	and origination year	and origination year	
Cluster	city	city	city-year
Observations	679,483	565,084	539,815
Pseudo R-squared	0.439	0.454	0.449

Note:

This table studies the robustness of our results in Table 3 in subsamples that allow better control for unobservables (e.g., location and condo market price dynamics). Column (1) repeats the analysis in Table 3 in the top 50 cities in the BBX sample with finer zip code fixed effects. Columns (2)-(3) perform the analysis in Table 3 using the merged Zillow-BBX subsample in the top 50 cities, where we replace the state-level HPI information with the zip code level values based on transaction prices obtained from Zillow, for both single family houses and condos. Specifically, *Log_Condo (Single family) Price Level* refers to the log of the monthly zip-level Zillow condo (single family) market average transaction price; *Condo (Single family) Price growth* is the monthly change in log zip-level Zillow condo (single family) price level. The dependent variable *D_default within 2* yrs takes the value of one for defaulting within two years of the loan origination date. We include the same set of control variables and similar fixed effects as in Table 3 (unreported here for brevity). Odds ratios are reported and robust z-statistics are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Condo loan default patterns are also tested in several “sand states” (i.e., California, Florida, Nevada, and Arizona), which exhibit more striking default patterns during the crisis. In unreported analyses, it can be confirmed that the condo loan default level and growth patterns are qualitatively the same among sand states as in the full sample.[31] Another potential sample selection bias could arise from a few super star cities whose condo markets have unique characteristics that could confound our interpretation. Robustness tests of the key default analysis (Table 3.2 and 3.3) are performed by removing New York and Los Angeles from our sample. The results remain qualitatively the same and are not reported in this chapter (but are available upon request).

3.5.2 Lender or borrower effect

The previous analysis suggests that the results in Table 3.2 and 3.3 are unlikely driven by location or asset market differences. However, it still remains an open question as to whether lenders or borrowers account for the observed

[31] They do not appear to be stronger in these sands states, likely because there are other important determinants of condo loan presence and growth (e.g., supply constraints and demographic distribution) that make the four states a crude and noisy identification (of cross-sectional heterogeneity).

default pattern difference. For example, different lenders may specialize in one particular asset market in loan origination, and there exists different screening standards across lenders (see, Rajan, Seru and Vig, 2013). As a result, the differential default patterns between the two markets reflect the fact that riskier borrowers are drawn to the condo loan market over time. Unfortunately, BBX does not contain lender information. The question is first approached in an indirect way. Using OLS, loan default is regressed on all the observables (as in Table 3.2 and 3.3) for condo and single family market separately for each cohort year. The R-square statistics are obtained for each of the 10 regressions, and compare the trend in R-squares between the condo market and the single family market (Figure 3.3). The rationale is as follows. The null hypothesis is that faster condo loan defaults over origination cohorts are due to the increasingly lax screening by some lenders who happen to originate more condo loans. Even though we are able to control for the observable differences in loan characteristics (e.g., riskier contracts among condo loans), lenders may select based on other unobservable information. If lenders apply different selection criteria over loan cohorts that are unobservable to econometricians, then the proportion of loan defaults unexplained by observed loan and borrower characteristics should increase in later origination cohorts, especially for condo loans. In other words, the null hypothesis implies a diverging R-square trend between the condo and single family loans. However, the pattern in Figure 3.4 reveals a similar trend in the R-squares in the two markets. In particular, the R-squares of the origination cohort 2006 and 2007 in two markets are observationally indistinguishable from each other. Since condo loan defaults peaked (relative to the single family loan defaults) in these two cohorts, these results provide the first piece of evidence that lenders and the associated time trend in credit supply may not be an important reason underlying the (differential) condo default pattern.

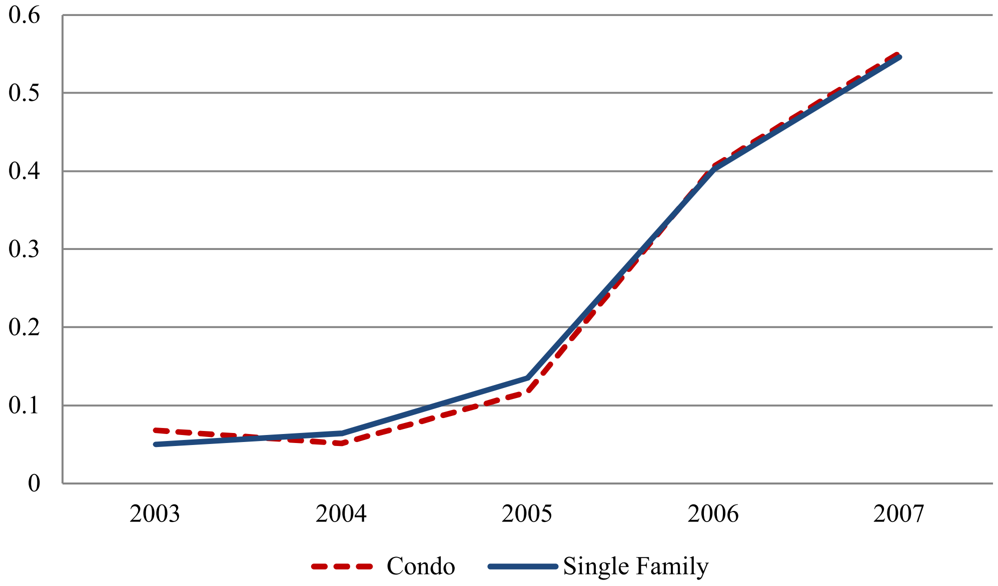


Figure 3.4 R square compassion from OLS regressions: condo and single family by year

Note:

This figure shows the trend in R-square statistics for condo and single family market separately for each loan origination year (over the period of 2003-2007). The R-squares are obtained from 10 (OLS) regressions, using the same independent and explanatory variables as in Table 3.2, by restricting to condo loans (or single family loans) within each origination year. The Y-axis indicates the R-square statistics, and the X-axis indicates the origination year of the loan.

3.5.3 Disentangling competing explanations: further evidence using Freddie Mac data

The previous analysis cannot completely eliminate concerns of a supply-driven channel. There is no lender information in the BBX sample, and the fact that many loan characteristics are (increasingly) riskier over time for condo loans presents another identification challenge.

In this section, another data source is introduced—loan-level performance data from Freddie Mac—to complement this analysis. Freddie Mac does not cover the subprime mortgage market, so the main analysis on the condo and the single market subprime market comparison and interaction is not feasible based on the Freddie Mac sample. However, it serves as a great supplementary dataset for the following reasons. Freddie Mac loan-level data contain lender identity information, which allows this study to better differentiate between the borrower channel and the lender-specific effect. Homogeneity among Freddie Mac loans (e.g. in contract terms) also facilitates a better identification against observed or unobserved heterogeneity in the condo loan market. The same filtering rule is applied to the Freddie Mac loan dataset, and the final Freddie Mac sample covers 3.79 million loans, out of which 11% are condo loans. Although smaller than the BBX sample, Freddie Mac's condo loan fraction is economically significant which ensures a meaningful comparison. The detailed summary statistics of the final Freddie Mac sample are shown in the Appendix.

Table 3.5 presents logistic regression results using the Freddie Mac sample. The available borrower and loan characteristics are included as control variables (e.g., FICO, LTV, owner occupancy status, and loan balance). Consistent with the analysis using BBX data, aggregate macroeconomic variables are included. All specifications include zip area, origination cohort, and year fixed effects, and standard errors are clustered at the zip area level.[32]

The baseline specification in Column (1) of Table 3.5 is closest to our BBX analysis in Table 3.2. It shows consistent results. On average, condo loans have a smaller likelihood of defaulting within two years after origination. However, the default rate in the condo loan grows at a far greater speed and eventually exceeds that in the single family market. The economic magnitude is comparable to that documented in Table 3.2: condo loans originated in 2007 are 10% (vs. 6.4% in Table 3.2) more likely to default than single family loans in the Freddie Mac sample. This result first provides external validity to our main analysis in Table 3.2 by using an independent data source. Furthermore, it sheds light on the interpretation of the observed default pattern in the condo loan market. Freddie Mac loans are homogenous: they are 30-year fully amortizing fixed rate mortgages for both condo and single family loans, in contrast to the prominent differences between the two types of loans in contract terms and subprime status in the BBX dataset. The observation that the default difference in the two markets is both qualitatively and quantitatively similar in a homogeneous sample of loans

[32] Freddie Mac only releases the location for each loan up to the first three digits of the exact zip code. We use this “zip area” as our location fixed effects for all analysis using the Freddie Mac sample.

complements the previous evidence by ruling out riskier contract terms or other unobserved heterogeneity among condo loans as the potential explanation.

In Column (2) of Table 3.5, lender information provided by Freddie Mac and lender fixed effects are included in the logistic regression. The results hardly differ from Column (1). A time-varying lender effect is allowed in Column (3) to control for a potential time trend in lending standard and results remain almost the same as in Column (1). In addition, the R-square improvement when we add lender-related fixed effects is negligible. This is consistent with the observation that the credit quality of approved condo borrowers (e.g., FICO score, LTV) does not deteriorate over time (Table A1, Panel B). While there still may be a common trend in credit supply that explains the default pattern for both the single family and condo loan market, the supply channel is not able to explain the faster growth trajectory among condo loans. Taken together, results in Column (2) and (3) add support to evidence in Figure 3.3: the lender or credit supply channel is unlikely an important factor in explaining the faster default growth pattern of condo loans.

Table 3.5 Controlling for Lender and Contract-term Differences: Evidence using Freddie Mac Data

	(1) D_default within 2yrs	(2) D_default within 2yrs	(3) D_default within 2yrs
D_Condo	0.439***	0.450***	0.459***
	(-8.01)	(-7.95)	(-7.84)
D_Condo * D_OrigYear2004	1.250*	1.239*	1.222*
	(1.95)	(1.88)	(1.75)
D_Condo * D_OrigYear2005	1.592***	1.625***	1.537***
	(3.68)	(3.89)	(3.39)
D_Condo * D_OrigYear2006	2.069***	2.061***	1.992***
	(5.94)	(5.99)	(5.79)
D_Condo * D_OrigYear2007	2.510***	2.455***	2.435***
	(6.41)	(6.48)	(6.60)
D_Owner occupied	0.914**	0.917**	0.925**
	(-2.29)	(-2.21)	(-2.02)
Original LTV	1.050***	1.049***	1.049***
	(41.94)	(41.00)	(41.24)
Log_FICO score	0.000***	0.000***	0.000***
	(-66.08)	(-66.66)	(-67.38)
Log_Original loan balance	0.752***	0.753***	0.755***
	(-8.41)	(-8.42)	(-8.32)
Log_Duration	0.923***	0.924***	0.923***
	(-30.76)	(-30.62)	(-29.97)
Log_HPI	0.026***	0.028***	0.030***
	(-12.96)	(-13.30)	(-13.15)
Unemployment rate	1.234***	1.233***	1.233***
	(10.52)	(10.91)	(11.23)
Yield slope	0.904***	0.904***	0.906***
	(-3.67)	(-3.74)	(-3.67)
Credit spread	1.401***	1.406***	1.412***
	(19.27)	(19.53)	(19.60)
Current interest rate	1.520***	1.522***	1.499***
	(14.66)	(15.00)	(14.56)
Fixed effects	zip-area, current year and origination year		
		Lender	Lender*origination year
Observations	1,823,656	1,823,656	1,823,656
Pseudo R-squared	0.354	0.358	0.360

Note:

This table presents the result of logistic regression analysis for the Freddie Mac full sample. This dataset includes only single-family and condominium (condo) purchase loans (< 10million USD loan origination amount) from all states originated during the period 2003–2007. Column (1) includes "zip-area" and current year fixed effects, as well as origination year fixed effects. Column (2) includes lender fixed effects, in addition to "zip-area" fixed effects, current year fixed effects, and origination year fixed effects. Column (3) includes the interaction of lender and origination year fixed effects, in addition to "zip-area" fixed effects, current year fixed effects, and origination year fixed effects. The dependent variable D_default within 2 yrs takes the value of one for defaulting within two years of the loan origination date. Please refer to Table A3 for definitions and summary statistics of the independent variables. Standard errors are clustered at "zip-area" level, and the fixed effects are not reported. Odds ratios are reported and robust z-statistics are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

3.5.4 Why are condo borrowers riskier?

It is shown, using different approaches on multiple datasets, that the default likelihood of condo loans increases at a greater speed and exceeds that of the single family loans for later origination cohorts, even after controlling differences in locations, asset market dynamics, contract terms and lending practice between the condo and the single family market. Therefore, riskier borrowers in the condo market emerge as the leading explanation. In this section, why borrowers in the condo market are riskier is investigated. Condo borrowers are unlikely the low credit quality borrowers who default because they cannot pay or refinance their mortgages as house prices start to decline. In the sample, compared to the single family market, condo borrowers have higher FICO scores, use subprime mortgages less frequently, and on average are charged a lower interest rate.

The recent literature highlights the role of investors in understanding defaults during the housing bust. Haughwout, et al. (2011) suggests that real estate investors rely on financial leverage in their purchases and default more ruthlessly when the housing market condition deteriorates. The price run-up in the earlier years of the decade attract more trend-chasing investors (e.g., Bayer et al., 2011; Fu and Qian, 2013), who are more likely to make a pure economic decision of deciding to default after a significant price drop. Real estate investors likely prefer condominium units due to their smaller size (and thus smaller investment), greater rental demand and less maintenance cost. Therefore, it is likely to observe a larger presence of investors in the condo borrower population over time. In the data of this study, condominiums are indeed more likely to be investment properties; a larger proportion of condominium purchases are for non-owner-occupancy purposes (31% in BBX and 22% in Freddie Mac) compared to single family purchases (27% in BBX and 10% in Freddie Mac). In addition, the share of owner-occupied loans decreased over time in both datasets. The investor channel is thus hypothesized to explain the observed default patterns in the condo market.

This study further examines whether investment-driven loans have a higher likelihood of default than non-investment-driven loans. The analysis is performed on the more homogeneous Freddie Mac sample that allows us to better control for heterogeneity among condo loans (e.g., in contract terms). In addition, investment-driven purchases can be identified with better precision in the Freddie Mac data: such information is incomplete and inaccurate in the BBX data where 41% of the loans have low or no documentation, compared to all Freddie Mac loans that have full documentation. Thus investment purchase dummy provided in the Freddie Mac data is used as our key independent variable in the analysis.

Panel A of Table 3.6 examines whether investment-associated condominium loans are associated with a higher likelihood default. The interaction between the condominium loan dummy and the investment purchase dummy in Column (1) supports the hypothesis: investment-associated condominium loans are on average 30.3% more likely to default than non-investment-associated condominium loans during the 2003-2007 origination period. Furthermore, within the condominium loan market, investment-associated loans' defaults grow at a much faster rate over origination cohorts (Column (2)). Importantly, while we show in Table 3.5 that condo loans of later cohorts (e.g., originated in 2007) on average are 10% more likely to default within two years than the same cohort single family loans, the higher default level among condo loans in that cohort is driven by investment-associated condo loans. Non-investment-associated condo loans originated in 2007 are 19.4% less likely to default than the same cohort single family loans, and investment-associated condo loans originated in 2007 are 88.7% more likely to default than the same cohort single family loans. This is strong evidence supporting the investor channel explanation, for the observed default pattern in the condominium market.

In Panel B of Table 3.6, the hypothesis is further tested by taking a closer look at the default behavior within the condo market. Again, the first column in Panel B suggests a strong investor effect: the investment-driven condo loan is 24% more likely to default within two years during our sample period, compared to condo loans not intended for investment purchases. This analysis seeks to further understand the investor channel by interacting the investment dummy with an *Option_to_default* variable that captures the moneyness of the default option (Column (2), Table 3.6).[33] Investors should be more responsive in their default

[33] Specifically, among the condo loan that have defaulted within two years in our sample, we define *Option_to_default* to be 1 if the current loan amount one month before default is greater than the average condo transaction prices (obtained from Zillow) in the same zip area during the same month. For those that have not defaulted within two years during our sample, *Option_to_default* is equal to 1 if, in at least one month during the first 24 months after origination, the loan amount in the current month is greater than the same-month average condo transaction price in the local area (i.e., =1 if borrower ever has one in-the-money default option during the first 2 year period after origination).

behavior when the current loan amount is greater than the value of the property. Results in Column (2) show that default likelihood significantly increases when the default option is in the money (as proxied by our *Option_to_default* dummy). Conditional on the default option being in the money, investment condos are 141% more likely to default within two years after origination than non-investment condos. On the other hand, when the default option is not in the money, there is no difference in default probability between the investment and non-investment condo loans.[34] Overall, the evidence suggests that investors play an important role in explaining condo defaults. In particular, condo investors (more) ruthlessly default when the current loan amount is greater than the property value.

[34] We also compare the risk profiles of investor condo loans with those of investor single family loans, using the same specification as in column 1 of Table VI, Panel B. In unreported results, we find no difference in the two-year default rates between investor loans in the condo market and investor loans in the single family market. This further suggests that borrower type (i.e., investors) is responsible for the pattern we identify in the paper.

Table 3.6 Investor Channel Analysis: Evidence from Freddie Mac

	(1) D_default within 2yrs	(2) D_default within 2yrs
Panel A: Full sample		
D_Condo	0.850** (-3.32)	0.452** (-7.92)
D_Condo*D_Investment	1.303** (3.06)	
D_Condo * D_OrigYear2004		0.791** (-3.50)
D_Condo * D_OrigYear2005		0.791** (-3.75)
D_Condo * D_OrigYear2006		0.752** (-3.88)
D_Condo * D_OrigYear2007		1.784** (6.68)
D_Condo* D_Investment* D_OrigYear2004		1.225 (1.68)
D_Condo* D_Investment* D_OrigYear2005		1.691** (4.20)
D_Condo* D_Investment* D_OrigYear2006		2.105** (6.12)
D_Condo* D_Investment* D_OrigYear2007		2.340** (6.18)
Fixed effects	zip-area, origination year, current year and lender	
Observations	1,823,656	1,823,656
Pseudo R-squared	0.357	0.358
Panel B: Condo market subsample		
D_Investment	1.242** (2.35)	0.936 (-0.56)
Option_to_Default		2.077*** (9.42)
Option_to_default*D_Investment		2.575*** (4.73)
Fixed effects	zip-area, origination year, current year and lender	
Observations	169,224	167,966
Pseudo R-squared	0.421	0.427

Note:

Panel A of this table presents the result of logistic regression analysis for the Freddie Mac full sample (condominium + single family loans) from all states originated during the period 2003-2007, and Panel B of this table shows the result of logistic regression analysis using only condominium (condo) purchase loans (< 10million USD loan origination amount) out of the full sample. The dependent variable *D_default within 2* yrs takes the value of one for defaulting within two years of the loan origination date. The definitions of the independent variables are same as in the summary statistics of the Freddie Mac sample (Table A3). *D_Investment* equals one if the use of the property if for investment (as recorded in Freddie Mac). Among the condo loans which default within 2 years, *Option_to_Default* is a dummy equal to one if the difference between current loan amount at one month before the defaulting month and Zillow zip-level condo HPI at the same month is larger than 0. For those condo loans which do not default within 2 years, *Option_to_Default* is equal to one if the difference between current loan amount and the current Zillow HPI is positive for at least one month during the first 24 months after origination. *Log_HPI* is log of the MSA-level quarterly FHFA/OFHEO House Price Index. Standard errors are clustered at “zip-area” level. “Zip-area”, origination cohort, year, and lender fixed effects are included but not reported. Odds ratios are reported and robust z-statistics are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

3.6 The Aggregate Implications of Condo Loan Defaults

Strong and robust evidence is documented that condo loans are inherently riskier than single-family loans. In particular, condo loan default rate grows at a fast rate and those loans in later origination cohorts default more even compared with subprime loans in the single-family market. Does the higher risk have aggregate implications for the recent housing crisis? The evidence suggests that investment-driven, riskier borrowers in the condo market are the most plausible driver for the observed default patterns in this market. Furthermore, investors are more responsive to market conditions in their default behavior. Given such, we conjecture that condo loans potentially default earlier than single family loans, and their earlier defaults potentially spill over by prompting more subsequent defaults in the single-family sector of the same geographic area.

This study examines this hypothesis in two steps. First, whether the within one year default likelihood among condo loans exhibit the same trend over time is studied. Second, whether early condo defaults predict subsequent defaults among single family loans with the same origination cohort located in the same local area is examined. This analysis focuses on the potential spillover effects to the subprime sector of the single family loan market. Since the implication of condo loan defaults on the single family subprime market is studied, this research uses the BBX sample for the analysis in the section.

3.6.1 Within-one-year default analysis

This analysis explicitly studies the within-one-year default decision, defined to be one if the loan is at least 60 days delinquent within the first year of loan origination. The one-year default probability of condo loans is compared with that of subprime loans in the single-family market (Table 3.7, Panel A). Similar as the finding in Table 3.3, condo loans' within-one-year default rate grows faster; in particular, condo loans' within-one-year default rate is greater than that in the single family subprime market for the later origination vintages. Condo loans originated in 2007 are 8.8% more likely to default within the first year of origination than single-family subprime loans of the same cohort. Next, the one-year default rate of condo and single family loans within the subprime sector are compared (Table 3.7, Panel B). Similar evidences are found. Particularly, within the subprime market, condo loans originated in 2007 are 87% more likely to default within one year of origination than single-family loans originated in 2007 (Column 2). Overall, the evidence in Table 3.7 is consistent with the argument that condo borrowers are more responsive to the market condition and experience more early defaults when the housing market condition deteriorates (i.e., in the later origination cohorts).

Table 3.7 Within-One-Year Default Analysis

Panel A: Logistic analysis of all condo loans and single-family subprime loans

	(1)	(2)
	D_default within 1yr	**D_default within 1yr**
D_Condo	0.819***	0.595***
	(-3.39)	(-2.84)
D_Condo * D_OrigYear2004		0.779
		(-1.21)
D_Condo * D_OrigYear2005		0.876
		(-0.75)
D_Condo * D_OrigYear2006		1.779***
		(3.24)
D_Condo * D_OrigYear2007		1.829***
		(3.10)
Fixed effects	City, current year and origination year	
Observations	626,419	626,419
Pseudo R-squared	0.445	0.447

Panel B: Logistic analysis of all condo subprime loans and single-family subprime loans

	(1)	(2)
	D_default within 1yr	**D_default within 1yr**
D_Condo	1.045	0.874
	(0.66)	(-0.47)
D_Condo * D_OrigYear2004		0.862
		(-0.40)
D_Condo * D_OrigYear2005		0.938
		(-0.20)
D_Condo * D_OrigYear2006		1.388
		(1.05)
D_Condo * D_OrigYear2007		2.140**
		(2.10)
Fixed effects	City, current year and origination year	
Observations	343,628	343,628
Pseudo R-squared	0.451	0.452

Note:

This table presents the results of the within-one-year default logistic regression analysis. The sample includes all condo loans and subprime loans originated during the period 2003–2007 in the BBX sample. Panel A presents the logistic regression results of all condo loans and single-family subprime loans, and Panel B presents results of condo subprime loans and single-family subprime loans. The dependent variable *D_default within 1 yr* takes a value of one for defaulting within one year of the loan origination date. We include the entire list of control variables; refer to Table 3.2 for the full regression list and Table 3.1 for the definitions of the variables. Standard errors are clustered at city level. City fixed effects, current year fixed effects and origination year fixed effects are included in the regression but not reported. Standard errors are clustered at the city level. Odds ratios are reported and t-statistics are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

3.6.2 Do (early) condo defaults predict the single family subprime market's subsequent default rate?

Next, we investigate the effect of condo loan defaults on the same-cohort single-family subprime loan market within the same zip code. Specifically, we study whether the one-year defaults of condo loans positively predicts second-year defaults of the same-cohort single family subprime loans in the same zip code.

From the BBX loan-level sample of all the condo and single-family subprime loans, the dependent variable$SF\ subprime\ \ 2nd\ year\ default\ (\%)_{j,t}$ is computed as the proportion of single-family subprime loans in the zip code j originated in year t that defaulted during the second year after origination. The main independent variable is $Condo\ within\ 1\ year\ default\ (\%)_{j,t}$, the proportion of condo loans in zip code j originated in year t that defaulted in the first year after origination.

To control for the within-subprime-market dynamics, $SF\ subprime\ within\ 1\ year\ default\ (\%)_{j,t}$ is included, the proportion of single family subprime loans in zip code j originated in year t that defaulted in the first year after origination, in the regression. Also, the MSA level HPI and the fraction of condo loans originated in the same year in the same zip code are controlled in the regression. Zip code fixed effects are included to allow any time-invariant location effects at the zip code level, and state-origination year fixed effects to control for any time-varying macroeconomic conditions at the state level. The standard error is clustered to allow correlation among zip codes within the same state in a given year.

In Column (1) of Table 3.8, we report the full sample result of regressing the proportion of the second-year defaults of single-family subprime loans originated in year *t* in zip code *j* on the proportion of first-year defaults of the same-cohort condo loans in the same zip code. Within the same cohort, a higher level of within-one-year defaults in the condo loan market positively predicts subsequent defaults of the single-family subprime loans in the second year after origination. (Unreported) granger causality tests verify the temporal lead-lag relationship: condo defaults precede the subprime mortgage defaults in the single family market. Intuitively, more of the condo loan borrowers are investors who are more likely to default strategically and at lower levels of negative equity. In addition, they may be less attached to the neighborhood (e.g., school district) and therefore more inclined to walk away earlier. This provides new insight on the triggering event of the housing crisis.

To increase the power of the test and to better identify the channel of the lead-lag effect, further analysis is performed. Specifically, the exogenous variation in state foreclosure laws is used to better identify the spillover effect of condo market defaults. Due to a faster foreclosure process, loan defaults in the non-judicial states lead to more foreclosures through which they have a greater impact on local house prices (Mian et al., 2012). Given this intuition, early condo loan defaults likely have a greater impact on the subsequent single family subprime loan defaults in the non-judicial states. The subsample analysis based on judicial and non-judicial states subsamples are reported in Columns (2) and (3) of Table 3.8. The within-one-year defaults in the condo market *only* positively and significantly predict second-year defaults among single-family subprime loans of the same origination cohort in the same zip code in the non-judicial foreclosure states. In judicial foreclosure states, the coefficient is insignificant.

Table 3.8 Do Early Condo Defaults Predict Subsequent Single Family Subprime Defaults?

	(1)	(2)	(3)
	SF subprime 2nd Year default (%) $_{j,t}$	SF subprime 2nd year default (%) $_{j,t}$	SF subprime 2nd year default (%) $_{j,t}$
	Full Sample	**Judicial Foreclosure States**	**Non-Judicial Foreclosure States**
SF subprime within 1 year default (%)$_{j,t}$	0.038**	0.026	0.044**
	(2.24)	(0.97)	(2.03)
Condo within 1 year default (%)$_{j,t}$	0.019	-0.024	0.033*
	(1.23)	(-1.38)	(1.80)
% Condo loans$_{j,t}$	-0.023***	-0.016**	-0.028***
	(-3.93)	(-2.19)	(-3.38)
Constant	0.041***	0.027***	0.032***
	(17.92)	(10.37)	(10.07)
Fixed effects		Zip code and state-origination year	
Cluster		state-origination year	
Observations	33,564	13,978	19,586
R-squared	0.406	0.344	0.421

Note:

This table reports the zip code level analysis of the single-family subprime market defaults from loans by their origination cohort years (2003–2007). From the loan-level sample with all the condo and single-family subprime loans, we compute $SF\ subprime\ \ 2nd\ year\ default\ (\%)_{j,t}$ as the proportion of single-family subprime loans in the zip code j originated in year *t* that default during the second year after origination. $SF\ subprime\ within\ 1\ Year\ default\ (\%)_{j,t}$ ($Condo\ within\ 1\ year\ default\ (\%)_{j,t}$) is defined as the proportion of single-family subprime loans (condo loans) in zip code j originated in year t that default in the first year after origination. % $Condo\ loans_{j,t}$ is the number of condo loans divided by the total number of single-family subprime and condo loans originated in year *t* in zip code *j.* We also include zip code and state-origination year fixed effects in all specifications and cluster the standard errors at state-origination year level. T-statistics are included in parentheses, and ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Next whether condo loan defaults predict subsequent house prices in the single family market is directly studied. If the economic intuition on the channel of the spillover effect is correct, one should observe a negative relationship between current loan defaults in the condo market and subsequent house price growth in the single family market within the same zip code. Using data from Zillow, we compute the (log) annual change in the zip code-level average transaction price in the single family market in year *t+1* to be our dependent variable. The key explanatory variable is the fraction of (two-year) loan defaults by condo borrowers in the same zip code in year *t.* This study includes, as our control variables, the condo loan share, single family house price level in year *t* at the same zip code, as well as zip code and state-year fixed effects. The standard errors are clustered at the state-year level to allow correlation among zip codes in the same state at a given year. Results provide consistent evidence (Table 2.9). A higher level of the current year's condo default rate is associated with a significant drop in the next year's single family house price growth in the same zip code. Furthermore, the negative association is stronger in the non-judicial states (coefficient = -0.083) than in the judicial states (coefficient = -0.053). Though the null hypothesis cannot be rejected that the two coefficients are statistically the same, it is observed that the difference is economically large: the predictability of condo defaults on single family house price growth is 50% larger in the non-judicial states than that in the judicial states.

Table 3.9 Do Condo Defaults Predict Single Family House Price Growth?

	(1)	(2)	(3)
	SF price growth (%)$_{j,t+}$	**SF price growth** (%)$_{j,t+}$	**SF price growth** (%)$_{j,t+}$
		Judicial	**Non-Judicial**
	Full Sample		
		foreclosure states	**foreclosure states**
Condo default (%)$_{j,t}$	-0.075***	-0.053*	-0.083**
	(-2.81)	(-1.79)	(-2.33)
SF subprime default (%)$_{j}$	-0.093***	-0.028	-0.128***
	(-3.41)	(-1.39)	(-3.47)
SF price$_{j,t}$	-0.000***	-0.000***	-0.000***
	(-6.52)	(-7.97)	(-4.01)
% **Condo Loans**$_{j,t}$	-0.001	-0.017	0.015
	(-0.11)	(-1.47)	(0.94)
Constant	0.281***	0.270***	0.216***
	(9.90)	(10.37)	(5.40)
Fixed effects		Zip code and state-origination year	
Cluster		state-origination year	
Observations	35,513	15,237	20,162
R-squared	0.814	0.831	0.803

Note:

This table reports the zip code level analysis of the predictability of condo loan defaults on subsequent house price growth in the single family market. The original loan sample includes all condo and single family subprime loans with origination years between 2003 and 2007 in the BBX sample. We obtain the final sample in this analysis by aggregating to the zip code level and merging with zipcode-level price information from Zillow. $SF\ price\ growth\ (\%)_{j,t+1}$ is calculated as the (log) change in the Zillow zip code-level average transaction price in the single family market in year t+1. $Condo\ default\ (\%)_{j,t}$ ($SF\ subprime\ default\ (\%)_{j,t}$) is defined as the fraction of (within-two-year) loan defaults by condo (single family) borrowers in the same zip code in year t. $SF\ price_{j,t}$ refers to Zillow single family house price level in year t at the same zip code. $\%\ Condo\ Loans_{j,t}$ is the number of condo loans divided by the total number of single-family subprime and condo loans originated in year *t* in zip code j. We also include zip code and state-origination year fixed effects in all specifications and cluster the standard errors at state-origination year level. T-statistics are included in parentheses, and ***, ** and * indicate 1%, 5% and 10% significance, respectively.

The observed difference in the predictive power of earlier condo defaults between the judicial and non-judicial foreclosure states may be associated with other unobservable factors that lead to higher default rates for both the condo and single family subprime loans. For example, in the non-judicial states that have stronger creditor rights, lenders may screen less as a result of which the average borrower is riskier. A cleaner identification to isolate the house price externality channel would be to focus on locations near the borders of adjacent states that only differ on their foreclosure laws. Unfortunately, in the data sample, there are very few condo loan originations at the state borders in general, making the analysis infeasible.[35] This is perhaps unsurprising given that condominiums are typically located in more urban areas. Nevertheless, I argue that the selection issue is unlikely driving our results for the following two reasons. First, in the data (BBX and Freddie Mac), the share of condo loans is similar among judicial and non-judicial states (17.09% vs. 13.36%), whereas the selection argument would imply a higher concentration of (risky) condo loans in the non-judicial states. Second, we compare the default pattern, for all loans, between the judicial and non-judicial states in a regression analysis and there is no evidence of systematic differences. These pieces of evidence, albeit suggestive, are indeed consistent with Mian, et al. (2012): there are no systematic differences between non-judicial and judicial states (including at the state borders) in default rates, house price growth, leverage, fraction subprime, income, unemployment rate, racial mix, poverty, or education.

Overall, the results in Tables 3.7-3.9 are consistent with the hypothesis that condo loan defaults have aggregate implications beyond the condo loan market itself. Condo borrowers are more responsive to housing market conditions, leading to a faster growth in their within-one-year default likelihoods, compared to the single family subprime loans. The early condo loan defaults predict a higher subsequent default rate of the same-cohort single-family subprime loans in the

[35] Using the same identification of state borders as in Mian, et al. (2012), the identified condo loans constitutes only 1.3% of loans near the state borders.

same zip code, primarily in non-judicial states with an efficient foreclosure process.[36] This suggests that the predictability largely works through the mechanism of house price externality. Indeed it is shown that condo loan defaults strongly predict future single family house prices in the same zip code, especially in the non-judicial states.

3.7 Summary

In this chapter an overlooked yet potentially important segment of the mortgage market—the condominium loan market—is identified, in understanding the recent financial crisis. The number of condominium loan origination has increased by 15-fold between 2001 and 2007. During this time period, condo loans accounted for 15% of all U.S. residential loan originations, rising from 9% in 2001 to 16% in 2007. Moreover, condo loan defaults grows at a faster rate than single family loan defaults, even after controlling for observed loan and borrower characteristics. For loans originated in year 2006, condo loans are 6.4% more likely to default than single family loans of the same cohort, and 12% more likely to default than subprime mortgages – presumably the riskier loans—in the single family market.

Despite the fact that condo asset and loan market differs considerably from the single family market, it is shown that the observed default pattern among condo loans is not explained away by observed and unobserved heterogeneity

[36] As a further robustness check, the results confirm that the within-one-year condo defaults granger-cause the subsequent same-cohort single family subprime loan defaults at the same location.

associated with condo loans (such as location, asset characteristics, or loan contract term or contract type differences). It is further observed that the results remain robust after we control for (time-varying) lender fixed effects, which suggest riskier condo borrowers as the main explanation for the faster default growth among condo loans.

Condo borrowers are unlikely the low credit quality borrowers who default because they cannot pay or refinance their mortgages as house prices start to decline; in our sample, condo borrowers have better creditworthiness than single family borrowers. The investor channel is hypothesized to play a more important role in our context: the price run-up in the earlier years of the decade attract more investors who are also more likely to make a pure economic decision of deciding to default soon after a significant price drop (Haughwout, et al., 2011). Given that real estate investors are more present in the condominium market, the observed default pattern in the condo loan market thus may be associated with the investor behavior. Consistent with the hypothesis, investment-purchase condo loans are found to be much more likely to default compared to other condo loans, and the effect is strengthened when the option to default is more in the money.

Lastly, the results reveal the effect of condo loan defaults on the subsequent subprime loan defaults in the single family market. Specifically, early condo defaults within the same zip code positively predict subsequent defaults by subprime mortgages of the same origination cohort in the single family market. In addition, condo loan defaults are negatively associated with subsequent single family house price growth in the same zip code. This result provides new insight of the triggering event of the housing crisis. Using exogenous variation in state foreclosure laws, it is confirmed that the predictive effects of condo loan defaults concentrate in judicial foreclosure states, consistent with the explanation that earlier condo loan foreclosures prompted more defaults among subprime mortgages within the same location by exerting downward price pressure on the neighborhood house prices.

Results in this chapter imply that condo loan market is an important channel to understand the cause and transmission mechanism of the recent financial crisis especially from the perspective of borrowers and investors' behavior. The findings that condo borrowers, especially investors, are riskier also suggest that lenders need to exercise more scrutiny in their lending practice in the condominium mortgage market. From a public policy point of view, it is found that simply requiring more skin-in-the-game regulations for lenders and lower LTV for the borrowers under the Dodd-Frank law is only a partial solution from avoiding a similar crisis in the future. The evidence provides the first step in studying the cause and aggregate implications of the condo loan defaults. Future research is required to understand the role of borrowers, especially investors in that market in fueling and potentially exacerbating the crisis.

Chapter 4 Foreclosure Concentration and the Exercise of Mortgage Default Options

4.1 Introduction

In recent years, the United States has experienced a nation-wide crisis in the mortgage market with unprecedented number of defaults and foreclosures. However, mortgage defaults and foreclosures were not evenly distributed across space. Miami, Las Vegas, Phoenix, Detroit and Los Angeles are the hard-hit metropolitan (metro) areas with intensive foreclosures. Other metros such as Seattle, Houston and Atlanta have much lower foreclosure rates. Also within cities, foreclosures are more concentrated in some neighborhoods than in others. For example, in Los Angeles, the foreclosure rate in zip code 90056 (Ladera Heights in South Los Angeles) is about forty times more than that in zip code 90403 (Santa Monica in West Los Angeles) in April 2014[37]. Questions arise as what socio-economic consequences those concentrated foreclosures bring to urban neighborhoods.

Existing research has found foreclosures generate externalities to urban neighborhoods. For example, they lower the values of nearby properties, increase neighborhood violent crimes and cause high property turnovers (see, e.g., Harding

[37] According to RealtyTrac, May 2014, Los Angeles County Real Estate Trends & Market Info. http://www.realtytrac.com/statsandtrends/foreclosuretrends/ca/los-angeles-county, foreclosure rate is defined as the number of foreclosures divided by the total number of housing units in the zip code.

et al., 2009; Immergluck and Smith, 2006a; Gerardi and Willen, 2009). In this study, a novel approach is taken to try to answer a new question, which is how concentrated foreclosures affect the default decision of mortgage borrowers in the surrounding area. The question addressed can be intuitively understood as how seeing foreclosure signs in one's neighborhood affects someone's likelihood of and attitude towards exercising her mortgage default option to enter into default.

On the one hand, intense foreclosures in a neighborhood bring about a signal to nearby borrowers that should they choose to default, similar to borrower's strategic default decision (Riddiough and Wyatt, 1994 and Guiso et al., 2013). On the other hand, foreclosure contagion can induce more defaults, either due to observational learning from seeing foreclosures in one's neighborhood (Agarwal et al., 2011), or due to the change of view that default is immoral or the ease of the stigma effect of default, or due to behavioral responses such as herding (Seiler, et al., 2012).

Therefore, the ultimate impact of neighborhood foreclosure concentration on borrowers' default decision is an empirical question, which is investigated in this study. With a massive dataset of individual mortgage loans from BlackBox Logic (BBX), the performance of individual loans can be tracked, to measure foreclosure intensity in each urban neighborhood, and further to estimate a model of mortgage borrowers' delinquency decision that incorporates neighborhood foreclosure concentration effect. Comparing to existing studies, this work takes a novel approach to not only assess the impact of neighborhood foreclosure concentration on individual borrower's delinquency probability but also estimate the impact of foreclosure concentration on the borrower's sensitivity to negative equity. The latter estimate measures to a certain extent the changing attitude of borrowers towards default option exercise.

In the main analysis, this research focuses on the Los Angeles-Long Beach-Santa Anna metropolitan statistical area (the Los Angeles MSA). The sample includes over 12,000 fixed-rate subprime mortgage loans that were originated between 1998 and 2008 and tracked through the first quarter of 2014. The results show that on average neighborhood foreclosure concentration enhances borrowers' default option exercise during the study period – borrowers are more willing to enter into default when there are intense foreclosures in the neighborhood. However, interestingly, the impact of foreclosure concentration varies in different regimes: before 2007, higher neighborhood foreclosure intensity is associated with reduced borrower sensitivity; entering into the crisis period (2007-2011), the impact turns from negative to positive; and post 2012, the impact becomes insignificant. Such variations are considered to reflect the balancing of the information effect and the contagion effect I discussed above. For example, during the crisis period, the foreclosure contagion effect might have been the dominant force and outweighed the information effect, so a positive net impact is observed.

The net impact of neighborhood foreclosure concentration on borrowers' sensitivity to negative equity also varies across different borrower groups. For example, the positive impact is significantly stronger among Asian borrowers than among non-Asian borrowers, but smaller among female borrowers than among male borrowers. There is a U-shape in the relation between neighborhood average FICO score and the impact of foreclosure concentration on borrowers' sensitivity to negative equity. Both very high and very low FICO neighborhoods see increased borrower sensitivity. Finally, lower income neighborhoods see stronger relation between foreclosure concentration and borrower sensitivity. These heterogeneities are also consistent with the notion that the balancing of the information effect and the contagion effect is likely to differ across different borrower groups.

The aforementioned results can be generalized to the whole state of California. And these results are robust to alternative house price index (HPI) and different measures of neighborhood foreclosure concentration.

Understanding how mortgage borrowers make their default decisions is critical to mortgage default risk management, pricing and underwriting. Traditional studies of borrower decision focus on mortgage borrowers' own socio-economic status such as the borrower's FICO score, income constraint, and equity position. Recently, some researchers have tried to place borrowers into social networks to understand their default decisions (see, e.g., Gangel et al., 2013; Guiso et al., 2013; Seiler et al., 2013). This study follows this line of thoughts. But different from existing studies that rely on simulated data or survey data, this study uses actual default data. The findings indicate that peer behavior has great influence on borrower's actual default choice. Therefore, default models should incorporate such network effects.

From a policy perspective, understanding the impact of foreclosure concentration on borrower's delinquency decision is also important. Delinquency is the first step of loan default, and foreclosure is usually the last step. Typically, large numbers of foreclosures follow the wave of delinquencies. Interestingly, what this research finds is that concentrated foreclosures can feedback onto borrower delinquency. Therefore, during the crisis, mortgage default can be self-enforcing in certain neighborhoods – increased delinquencies lead to more foreclosures, and concentrated foreclosures further lead to even more delinquencies. From this perspective, timely intervention by the government to reduce foreclosure is important to break the loop and to stop the foreclosure cascade.

The rest of the paper is organized as follows: in the next section, the data and some results regarding neighborhood foreclosure concentration are presented in Section 4.2. The hypothesis development and model are discussed in Section 4.3, and empirical results regarding the impact of foreclosure concentration on borrower delinquency are shown in Section 3.4. Concluding remarks are in the final section.

4.2 Data and Measures of Foreclosure Concentration

4.2.1 Data Sources

The first and main data in this study comes from the loan-level data furnished by BlackBox Logic (hereafter BBX). The BBX aggregates data from mortgage servicing companies. The most recent BBX data contains roughly 22 million non-agency (including jumbo, Alt-A, and subprime) mortgage loans throughout the United States, making it a comprehensive source for mortgage default studies[38]. BBX provides detailed information on the borrower and the loan at loan origination, including the borrower's FICO score, original loan balance, interest rate, loan term (30 year, 15 year, etc.), loan type (fixed-rate, 5-1 ARM, etc.), loan purpose (home purchase, rate/term refinance, cash out refinance), occupancy status, prepayment penalty indicator and other characteristics. BBX also tracks the performance (default, prepayment, mature, or current) of each loan in every month.

Another key data source is the Home Mortgage Disclosure Act (HMDA) implemented by the Federal Reserve Board, which requires that lending institutions report virtually all mortgage application and origination data. HMDA is considered the most comprehensive source of mortgage data, covering about 80 percent of all home loans nationwide and an even higher share of loans originated in metropolitan statistical areas (Avery et al., 2007). In particular, it provides a nearly complete universe of 122 million U.S. mortgage applications over the

[38] The BBX data is comparable to other well-known datasets such as the CoreLogic data.

period 2001–2010. The key reason for using HMDA is that it covers borrower characteristics such as applicant's race, sex, and annual income that are not contained in the BBX data. HMDA also provides abundant information on the loan characteristics at the stage of loan application, including loan amount (in thousands), loan purpose (home purchase or refinancing or home improvement), borrower-reported occupancy status (owner-occupied or investment), (in the case of originated loans) whether the loan was sold to the secondary market within the year of origination, and other characteristics. Property-related variables available in HMDA are geographic location (census tract level identification) and property type (one-to-four-family or manufactured housing or multifamily).

Given the existence of common variables in the BBX data and the HMDA data, BBX loan-level data are matched with selected HMDA loan data using step-by-step criteria.[39][40] First, BBX loans are matched to HMDA loans with the same loan purpose and occupancy status of the borrowers. Second, based on the origination dates of BBX loans, HMDA loans within the same year of origination are considered. In addition, BBX loans are only matched to HMDA loans with the same zip code. Last, loans from BBX should have the same original loan amount as those from HMDA. For all possible HMDA matches for each BBX loan (with the same BBX identifier but different HMDA identifiers), only the first record for the same BBX identifier is kept. Any BBX loan that has no corresponding HMDA

[39] There is no unique common identifier of a loan from these two databases.

[40] In order to match with BBX data, only loan applications marked as originated in HMDA data are considered. Those loans originated by FNMA, GNMA, FHLMC and FAMC are removed. Those with loan type of FSA (Farm Service Agency) or RHS (Rural Housing Service) are excluded as well.

loans matched using the above criteria is a non-match and is excluded from our sample[41].

Furthermore, the loan-level data is merged with macro variables such as the MSA-level unemployment rate from Bureau of Labor Statistics, the CoreLogic Case-Shiller zip code-level Home Price Index, and the S&P Case-Shiller MSA-level Home Price Index. Treasury bond rate and interest rate swap rate from the Federal Reserve and mortgage interest rate from Freddie Mac are also matched into the data.

For the main tests, this research focuses on first-lien, fixed-rate subprime mortgage loans for the Los Angeles-Long Beach-Santa Anna metropolitan statistical area (the Los Angeles MSA).[42] The advantage of focusing on one particular MSA rather than pooling MSAs is that this analysis can be insulated from the cross-MSA disparities in borrower behavior that is due to legal and institutional differences, and thus gain cleaner inference from our model. Later our analysis is generalized to the whole state of California. This study focuses on the subprime mortgage loan sample that contains enough number of delinquencies, which enable us to estimate a sensible delinquency model.

[41] The success rate of our match is about 70 percent.

[42] A series of filters is also applied: we first exclude loans originated before 1998 for accuracy consideration; we also exclude those loans with interest only periods or those not in metropolitan areas (MSAs); loan occupancy status indicated as second home or vacancy home, loans with missing or wrong information on loan origination date, original loan balance, property type, refinance indicator, occupancy status, FICO score, loan-to-value ratio (LTV), documentation level or mortgage note rate are excluded.

4.2.2 Measures of Foreclosure Concentration

A number of neighborhood foreclosure concentrations are created measures at the zip-code level. The main measure is the foreclosure intensity calculated as the total number of foreclosures in the past two quarters (e.g. for 2009Q1 it is 2008Q4 and 2008Q3) divided by the total number of housing units (in thousands) in each zip code. A foreclosure intensity rank order of all zip codes in the Los Angeles MSA in each quarter is further created and then a dummy variable “High foreclosure intensity” is defined as the zip-quarter that ranks in the 90th percentile of all zip-quarters.

Alternative foreclosure intensity is also calculated as the total number of foreclosures in the recent four quarters (current quarter plus the past three quarters) divided by the total number of housing units in each zip code (in thousands). Accordingly, a “High foreclosure intensity” dummy variable is created based on rank order. Finally, instead of using the total housing units as the denominator to calculate foreclosure intensity, the total population in each zip code is used to calculate per capita foreclosure intensity measures.

Figure 4.1 shows some maps of foreclosure intensity. The first map shows the aggregate foreclosure intensity from all years. It is observed that there is great variation in foreclosure concentration across neighborhoods. Generally, zip codes in the inland cities have greater foreclosure intensity than those along the coast; zip codes in northern cities experience greater foreclosure intensity than those in west and east cities. Among all cities in this metropolitan area, Santa Clarita

Valley, Antelope Valley and San Fernando Valley experienced the highest foreclosure intensity: for every one thousand housing units in these zip codes, 50-135 loans during the period of 1998-2008 turned into foreclosures. San Gabriel Valley and Gateway Cities also suffered great waves of foreclosure during this period, ranging from 10 to 50 foreclosures per thousand housing units per zip code. Westside Cities, located in the west of this area, are shown to have the least foreclosure concentrations, with less than 10 foreclosures for every thousand housing units at the zip code level.

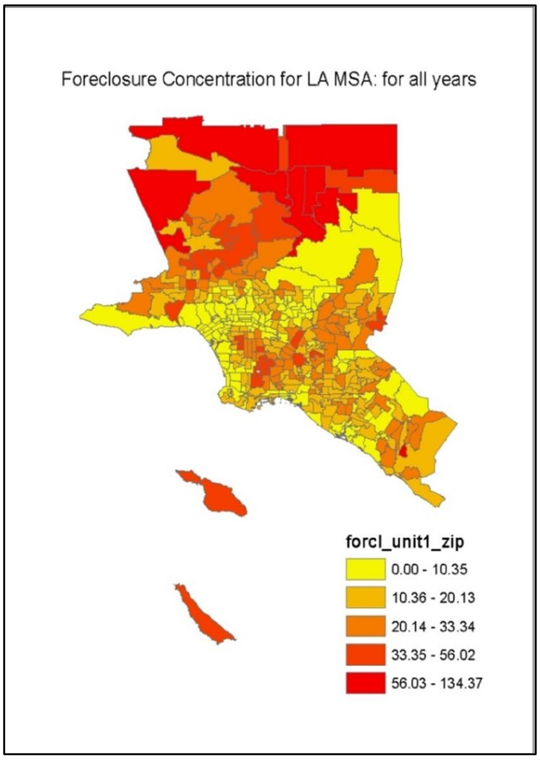


Figure 4.1 Foreclosure Concentration for LA MSA: for all years

Note:

Foreclosure concentration is calculated as the total number of foreclosures in the past two quarters (e.g. for 2009Q1 it is 2008Q4 and 2008Q3) divided by the total number of housing units (in thousands) in each zip code.

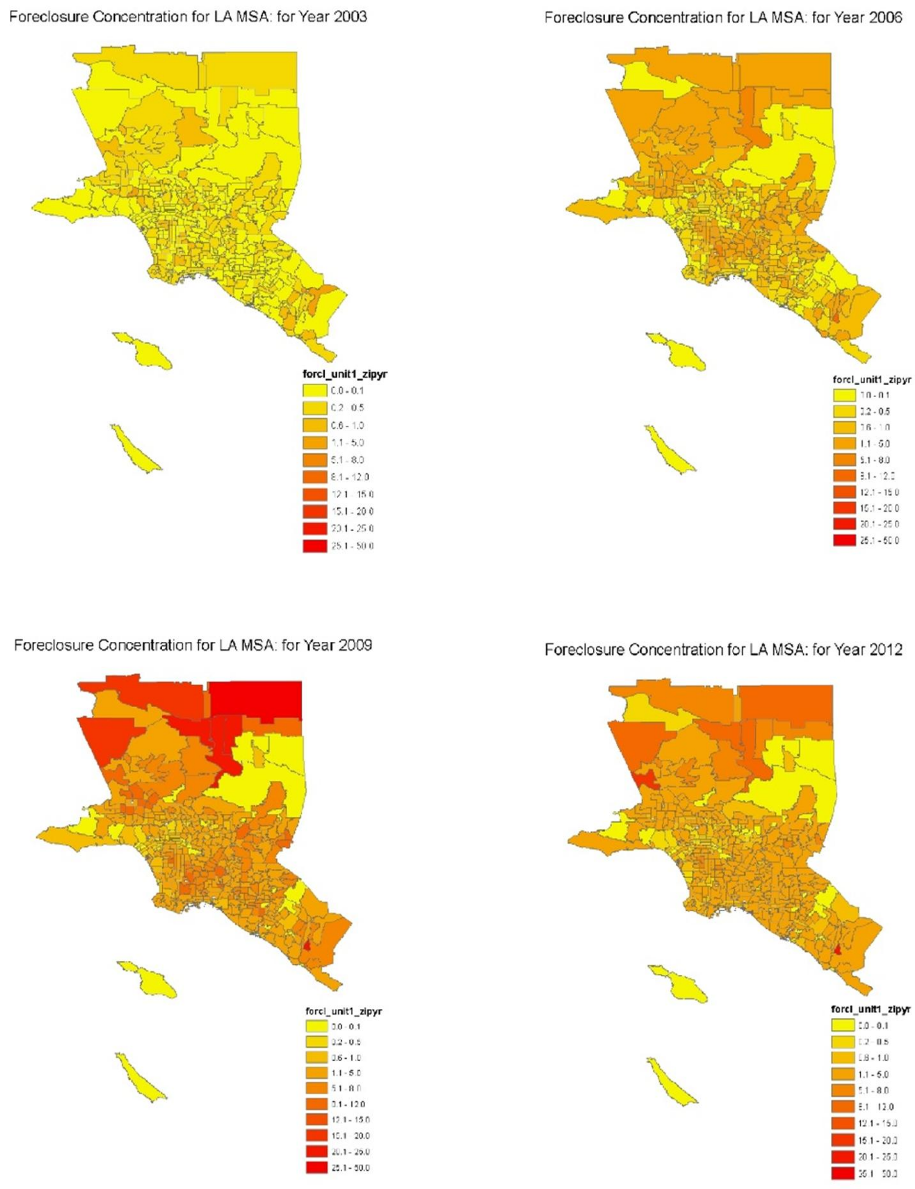


Figure 4.2 Foreclosure Concentration for LA MSA: for Year 2003, 2006, 2009 and 2012

Note:

Foreclosure concentration is calculated as the total number of foreclosures in the past two quarters (e.g. for 2009Q1 it is 2008Q4 and 2008Q3) divided by the total number of housing units (in thousands) in each zip code.

The foreclosure intensity maps are further created for each individual year and show the years of 2003, 2006, 2009 and 2012 in Figure 4.1. It is shown that foreclosure intensities vary significantly across the four years. 2003 and 2006 overall have small average foreclosure intensities (0.11 and 0.75 foreclosures per thousand housing units, respectively), in contrast to 3.05 and 1.81 foreclosures per thousand housing units in 2009 and 2012. Specifically, in 2003, more than half of the metropolitan area has foreclosure intensity of less than 1 per thousand, while in 2006, zipcodes in northern and southern cities experienced high foreclosure intensities, reaching 5 to 8 per thousand. In 2009, we observe even greater increases in the foreclosure concentration in the north, south and central area, with the greatest concentration in the northern part (25-50 per thousand). The foreclosure intensity in 2012, although still at a pretty high level compared with that in 2003 and 2006, starts to decrease, with the highest intensity of 20 per thousand housing units at the zip code level. The possible explanation for these phenomena is that the strong house price appreciation during 2003-2005 helped most of the loans in 2003 and 2006 out of foreclosure troubles, while the sharp and far-reaching house price decline starting from 2006 led to the much higher foreclosure concentration later in 2009 and 2012. The gradually recovering housing market in 2012 helped to reduce foreclosures. Among all cities in this area, Antelope Valley from the northern part experienced the most serious foreclosure problems through the four years, while San Gabriel Valley, the city located in the east of this area, remains at very low foreclosure intensity across the whole study period.

4.2.3 Descriptive statistics

Before digging into the main analyses, a preliminary look at the sample is taken here. Table 4.1 reports the number of originated loans in our sample by vintage, and Table 4.2 presents the numbers of loans at loan termination by the choice of default, prepay or current (censor).[43] As shown in Table 4.1, the number of loans originated rise slowly from 1998 (105 loans) to 2002 (512 loans), while starting from 2003 through 2006, there is a sharp jump in the observation numbers, with the highest number in 2005 (3,719 loans, 26% of the total sample) and lowest in 2003 (1,848 loans, 15% of the total sample). Since 2007, the loan number has declined quickly, with only about 61 percent less than that in 2006. The origination year distribution of our sample reflects the development of the subprime mortgage market. By looking at the loan numbers by termination status in Table 4.2, it is shown that among 12,007 loans in the sample, around 39% of loans have been defaulted, around 42% have been prepaid, and only 19% remain current by the time of January 2014.

[43] The terminations status of a loan is classified into default, prepay, and censor, whichever is the earliest at the end of January 2014. Default is defined as over 60- day delinquency. Prepay refers to early repayment of a loan, often as a result of refinancing to take advantage of lower interest rates. Current (censor) means that the loan is alive at the end of January 2014.

Table 4.1 Number of Loans in Our Sample by Vintage

Origination Year	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1998	105	0.87	105	0.87
1999	123	1.02	228	1.9
2000	184	1.53	412	3.43
2001	245	2.04	657	5.47
2002	512	4.26	1169	9.74
2003	1848	15.39	3017	25.13
2004	2625	21.86	5642	46.99
2005	3179	26.48	8821	73.47
2006	2290	19.07	11111	92.54
2007	895	7.45	12006	99.99
2008	1	0.01	12007	100

Note:

This table shows the frequency distribution of loan originations in our sample. We include first-lien, fixed-rate subprime mortgage loans for the Los Angeles-Long Beach-Santa Anna metropolitan statistical area (the Los Angeles MSA), and exclude those loans with interest only periods or those with missing or wrong information. All the loans are originated during the period 1998—2008.

Table 4.2 Performance of Loans in Our Sample

Termination type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Current	2245	18.7	2245	18.7
Prepay	5118	42.63	7363	61.32
Default	4644	38.68	12007	100

Note:

This table presents the frequency distribution of loan termination status in our sample, by the choice of default, prepay or current (censor), whichever is the earliest at the end of January 2014. Default is defined as over 60- day delinquency. Prepay refers to early repayment of a loan, often as a result of refinancing to take advantage of lower interest rates. Current (censor) means that the loan is alive at the data collection point—January 2014.

Table 4.3.1 reports the frequencies of some loan and borrower characteristics of our subprime FRM sample. Although approximately 52% of loans have full documentation of income, asset or employment, there are 26% of loans with low or even no documentation. Among the 12,007 loans in our sample, 10,394 loans (87%) have origination LTV greater than 80. Only 5% of loans are 15-year FRMs, which are usually thought to be less risky than 30-year FRMs. 97% of loans are classified as owner-occupied, compared with investment purpose (3%). Regarding property type, single family group ranks first (around 89% of the total loans), followed by condominium group. In terms of loan purpose, cash-out refinance and rate/term refinance account for about 94% of the total loans, while purchase loans only account for 6%. Consistent with the usual characteristics of subprime loans, about 87% of loans in the sample have prepayment penalty clause in the mortgage contracts, which might limit the subprime borrower's ability to refinance into more affordable loans and thus increase the chance of default. In terms of borrower characteristics, White and African American borrowers take up 49% and 12% of the total sample, while Asian borrowers are only 6%. More than 60% of the borrowers are male borrowers.

Table 4.3.2 shows the descriptive statistics of important loan and borrower characteristics. Because of high housing costs in Los Angeles MSA, our loans had an average original loan balance of $263,130. The average FICO score is 582, and the current interest rate reaches 7.22 on average. Borrower's debt-to-income ratio is around 28 percent on average. The average original LTV and combined LTV are both around 65 percent.

Table 4.3 Summary Statistics of Loan and Borrower Characteristics

Table 4.3.1 Frequency Distribution of Loan and Borrower Characteristics

		Frequency	Percent	Cum. Freq.	Cum. Pct.
Documentation type	Full doc	6245	52.01	6245	52.01
	Low doc	3028	25.22	9273	77.23
	No doc	147	1.22	9420	78.45
	Reduced doc	143	1.19	9563	79.65
	Unknown doc	2444	20.35	12007	100
LTV greater than 80 percent	Yes	10394	86.57	10394	86.57
	No	1613	13.43	12007	100
Race	White	5831	48.56	9147	48.56
	Asian	684	5.7	930	54.26
	Black	1430	11.91	2360	66.17
	Other	4062	33.83	12007	100
Gender	Male	7315	60.92	7315	60.92
	Female	4089	34.06	11404	94.98
	Unknown information	603	5.02	12007	100
Loan type	30-year FRM	11358	94.59	11358	94.59
	15-year FRM	649	5.41	12007	100
Property type	Single family	10631	88.54	10631	88.54
	PUD	341	2.84	10972	91.38
	Condo	1035	8.62	12007	100
Loan purpose	Home purchase	725	6.04	725	6.04
	Rate/term refinance	2142	17.84	2867	23.88
	Cash-out refinance	9140	76.12	12007	100
Occupancy status	Owner-occupied	11611	96.7	11611	96.7
	Investment property	396	3.3	12007	100
Prepayment penalty type	No	116	0.97	116	0.97
	Yes	10396	86.58	10512	87.55
	Unknown	1495	12.45	12007	100
Loan with a second lien	No	10043	83.64	10043	83.64
	Yes	1964	16.36	12007	100
Total number of loans			12,007		

Table 4.3.2 Means, Standard and Deviations of Loan and Borrower Characteristics

Variable	Mean	Std. Dev.	Minimum	Maximum
Original loan amount	263,130	131,013	22,000	2,500,000
FICO SCORE	582	34	417	804
Current interest rate (%)	7.22	1.11	1.64	13.83
LTV (%)	65	17.17	6	139
Combined LTV (%)	65	18	6	125
Payment-to-income ratio	0.28	0.16	0.01	11.37
Total number of loans		12,007		

Note:

Table 4.3 reports the summary statistics of loan and borrower characteristics in our sample. Table 4.3.1 presents the frequency distribution of some important loan and borrower characteristics, while Table 4.3.2 shows the mean, standard deviations, minimum and maximum of some numerical variables. Documentation type is an indicator whether a particular loan has full, low, do or reduced documentation of income, asset or employment. LTV greater than 80 percent is equal to Yes if the original loan-to-value (LTV) ratio is greater than 80 percent. Race refers to the racial group that the borrower belongs to, and Gender indicates whether the borrower is a male or female. Loan type means whether the durations of the FRM loan is 30 years or 15 years. Property type refers to the classification of the property securing the mortgage: i.e. Single family, PUD (planned urban development) and Condo (condominium). Loan purpose indicates the primary reason the mortgage was taken out by the borrower. Occupancy status means the use of the home such as investment, owner-occupied (primary residence), etc. Prepayment penalty type is an indicator denoting that a fee will be charged to the borrower if they elect to make unscheduled principal payments. Loan with a second lien is Yes if a second mortgage is taken out on the same property. Original loan amount is defined as the amount of principal on the closing date of the mortgage. FICO SCORE refers to the FICO (formerly the Fair Isaac Corporation) borrower credit score at the time of the loan closing. Current interest rate refers to the coupon rate charged to the borrower for the most recent remittance period. LTV (%) refers to the ratio of the original loan amount to the property value at loan origination, while Combined LTV (%) means the ratio of all loan amounts on the property at the time of origination to the property value at loan origination. Payment-to-income ratio refers to the percentage of monthly mortgage payment to borrower's monthly income.

4.3 Hypotheses and Methodology

4.3.1 Hypothesis Development

The hypothesis development in this section starts with a brief explanation of the default process. Typically borrower's failure to make monthly mortgage payment constitutes a default[44], which can result in a sale of the collateral to fulfill the borrower's debt obligation. However, default is not a one-stage process. It is

[44] Technically, borrower's failure to pay taxes or insurance premiums, failure to keep the property in repair, or violations of other loan covenants can also lead to default.

actually a multiple-stage lengthy process. The borrower first decides whether to miss a scheduled monthly payment. If a payment is missed and the loan becomes 30-day delinquent, late fees will be charged. Subsequently when a mortgage loan is 60-day overdue, a notice of default (NOD) is usually sent to the borrower, and the servicing of the loan will be transferred from the general servicer to a special servicer, who will first seek a workout if appropriate. If a workout is unsuccessful, the lender (through special servicer) will start the foreclosure process, which typically occurs after the loan is over 90-day delinquency. The actual foreclosure sale (trustee sale in non-judicial foreclosure states like California) typically takes another several months to occur because foreclosure has to be publicized fully (e.g., notification sent to the borrower, notice published at the local newspaper, and signs to be put on the property). Finally, if a sale is successful, the lender receives sales proceeds net of all the fees and legal costs. An unsuccessful foreclosure/trustee sale leads to real estate owned (REO), in which the lender obtains the title of the property. Therefore, we can see that borrower delinquency is the beginning of the default process while foreclosure is in the subsequent stage of default, which can be many months away down the road.

Many believe that mortgage borrowers are strategic in their delinquency decisions in a sense that they not only consider their ability to make the monthly payment, the current equity position (whether the house is worth more than the remaining loan balance) and house price trend (the possibility of a future recovery from the current negative equity position) but also consider what the lenders' reactions are. The logic is as follows: given that foreclosure is costly to lenders as sale proceeds from a foreclosure sale usually fall short of the remaining balance plus all the transaction costs, lenders usually first seek to “workout” a delinquent loan. A loan workout can take the form of a reduced interest rate/payment, reduction in loan principal, and extension of the mortgage term, which are typically in the borrower’s favor. Foreclosure is typically the worst outcome not only to the lender but also to the borrower, because it causes the borrower to lose her home and incur significant credit impairment. Therefore, from a game-theoretic perspective, borrower’s delinquency decision depends upon her own strategic perspective on the consequential gains or losses from acceptance, rejection, or a counter-offer from the lender and the likelihood of each response. Riddiough and Wyatt (1994) argue that borrower’s delinquency decision depends on how tough the lender is. Guiso et al. (2013) also argue that borrower’s altitude towards strategic default depends on her assessment of the probability of getting sued by the lender (a foreclosure).[45]

Following this line of thoughts, I would expect that the incidences of foreclosure, especially large number of foreclosures in one’s neighborhood can serve as a signal to the borrower that the chance of receiving a favorable loan modification or short sale is low while the chance of being foreclosed is high should she chooses to enter into default. Therefore, foreclosure concentration will discourage the borrower’s choice of delinquency. We define this effect as an information effect.

[45] Strategic default is when the borrower is able to make the monthly payment but chooses not to do so in anticipation of a favorable loan modification after she is delinquent on her loan.

On the other hand, recently there has been a growing literature on foreclosure contagion. Several studies have found that nearby foreclosed properties lower the price of neighboring properties (see, e.g., Immergluck and Smith, 2006b, Harding et al., 2008; Schuetz, et al., 2008; Campbell et al., 2009; Lin, et al., 2009). Although the exact mechanism of such foreclosure contagion is still debated in the literature, a compelling explanation is the observational learning suggested by Agarwal et al. (2012): homeowners update their beliefs about the value of their homes when they receive signals about house price trend. Foreclosures in one's neighborhood send out a public signal of a declining property market. Based on such a signal, nearby homeowners will adjust their valuation downward, causing an observed negative impact of nearby foreclosure on property values. Such downward adjustment in valuation apparently increase the probability of default as borrowers default their mortgage loans mainly because the value of the property is lower than the mortgage loan balance.[46] Therefore, from this perspective, concentrated foreclosure in one's neighborhood has a positive impact on someone's default decision.

The impact of concentrated foreclosures on borrower's delinquency decision can also arise from herding. People do not always exercise independent judgment due to social influence (Shiller, 1995). Meanwhile, in situations where information is limited individuals can follow the herd in the hope of gaining the superior information of the group (Bikhchandani et al., 1998). For these reasons,

[46] Some researchers argue that insolvency (e.g., loss of income) also cause default. However, if there is positive equity, the borrower should be able to sell the property and payoff the loan to avoid a default. Therefore, negative equity is the ultimate driver of residential mortgage default.

herd behavior can be a source of mispricing and speculative bubbles (Shiller, 2008). In a recent study, Seiler et al. (2014) find that homeowners are easily persuaded to follow the herd to strategically default their mortgage loan. Extending this herding rational to mortgage borrower's delinquency decision, I would expect someone who resides in a neighborhood with concentrated foreclosures is exposed to the influence of her neighbors and thus is more likely to exercise her default option when she sees many foreclosure signs in her neighborhood.

During the recent mortgage market crisis, there have been heated debates regarding whether it is immoral to default one's mortgage loan (see, e.g., White, 2010; Guiso et al., 2010). Although many Americans think it is immoral to strategically default their mortgage loan, seeing many neighbors have done so might have changed some borrowers' view. In addition, the thought that "I am not doing this alone" can ease the stigma effect of mortgage default and thus cause borrowers to be more willing to enter into default.

In summary, this study hypothesizes that foreclosure concentration can have both positive and negative impacts on borrower's delinquency decision. It is really an empirical question as to what the net impact is. Further, one may observe different net impacts in different times and across different borrower groups, depending on how those positive and negative impacts play out differently over time and in the cross section.

4.3.2 Methodology

In order to empirically assess the impact of foreclosure concentration on borrower's delinquency decision, we estimate a Cox proportional hazard model of mortgage delinquency. The hazard model is widely used in the mortgage literature (see, e.g. Vandell, 1993; Quigley and Van Order, 1995; An, et al., 2012). It is convenient mainly because it allows us to work with the full sample of loans despite some observations being censored when the data is collected. This is an important feature for this study because a large portion of the mortgage loan observations is censored.

Assume the hazard rate of default of a mortgage loan at period T since its origination follows the form

$$h_i(T, Z'_{i,t}) = h_0(T)\exp(Z'_{i,t}\beta). \tag{4.1}$$

Here $\mathrm{h}_0(\mathrm{T})$ is the baseline hazard function, which only depends on the age (duration), T, of the loan and is an arbitrary function that allows for a flexible default pattern over time47; $\mathrm{Z}'_{\mathrm{i,t}}$ is a vector of covariates for individual loan i that include all the identifiable risk factors. In this proportional hazard model, changes in covariates shift the hazard rate proportionally without otherwise affecting the duration pattern of default. Commonly used covariates include negative equity, FICO score, loan balance, the loan-to-value (LTV) ratio, payment to income ratio, and change in MSA-level unemployment rate.

[47] Notice that the loan duration time *T* is different from the natural time t, which allows identification of the model.

Neighborhood foreclosure concentration is the key variable that will be on the right hand side of the delinquency model. However, different from existing studies, we take a novel approach to not only include the measure of foreclosure concentration in this study as a covariate but also interact the foreclosure concentration measure with negative equity. In so doing, the coefficient of negative equity is allowed to depend on the measure of foreclosure concentration. The model estimated is thus

$$h_i(T, Z'_{i,t}) = h_0(T)\exp(Z'_{i,t}\beta)$$

$$Z'_{i,t}\beta = \beta_1 NegEq_{i,t} + \beta_2 ForclRate_{j,t} + \beta_3 ForclRate_{j,t} \cdot NegEq_{i,t} + X'_{i,t}\gamma \tag{4.2}$$

where $\mathrm{NegEq}_{\mathrm{i,t}}$ is negative equity of loan i in zipcode j at time t, $\mathrm{ForclRate}_{\mathrm{j,t}}$is the neighborhood foreclosure rate of zipcode j at time t, and $\mathrm{X}'_{\mathrm{i,t}}$ are other control variables such as FICO score, LTV ratio, etc.

Existing studies have found negative equity to be a critical driver of mortgage borrowers' default option exercise (see, e.g., Campbell and Dietrich, 1983; Quigley and Van Order, 1995; Deng et al., 2000). However, existing research has also found that mortgage borrowers do not always default when facing negative equity (see, e.g., Vandell, 1995; Deng and Quigley, 2002; Foote, et al., 2008). Therefore, the coefficient of negative equity in a delinquency model measures the sensitivity or responsiveness of the borrower to negative equity in her choice of delinquency. It can also be viewed as the borrower's attitude to exercise default option. Therefore, β_3 in equation (4.2) measures the impact of foreclosure concentration on borrowers' attitude to exercise their mortgage default option. Note that by including neighborhood foreclosure rate as a covariate, the direct impact of foreclosure concentration on delinquency probability is also measured (the impact is reflected in $\beta_2$in our model). In addition, this variable will control for any unobservable neighborhood characteristics that are orthogonal to house price movement and other measured changes in the neighborhood if there is any such unobservable characteristics.

4.4 Empirical Analysis on the Impact of Foreclosure Concentration on Borrower Delinquency

4.4.1 The impact of foreclosure concentration on borrower default option

The first set of estimation results is reported in Table 4.4. Model 1 is the model without time-fixed effect. In addition to the focus variables seen in equation (4.2), we have 25 control variables (the X variables). Most of these control variables are significant with signs conforming to existing research or economic theory. For example, low or no doc loans have higher risk of delinquency and borrowers of those loans are more sensitive to negative equity. Owner-occupied loans are less sensitive to negative equity than investor loans. FICO score is negatively correlated with delinquency probability but the function is concave. In addition, higher FICO score borrowers are more sensitive to negative equity. Large loans and loans with high LTV (over 80 percent) are more likely to enter into default, everything else equal. Rate/term refinance loans are less likely to be delinquent while cash-out refinance loans are more likely to become delinquent. Loans with higher payment to income ratio have higher delinquency risk, and African American borrowers and female borrowers are more likely to enter into default. Finally, increase in MSA level unemployment rate causes more delinquency.

Next the focus variables in this study are discussed. Consistent with findings in the existing literature, negative equity is a highly significant factor of mortgage delinquency. The higher the negative equity is the more likely the loan will be delinquent (the positive β_1). In addition, it can be seen from the significant positive coefficient of the square term of negative equity that the function is convex, which is as expected – borrowers become extremely sensitive when they have a large negative equity. The more interesting findings here are on the zip-level foreclosure rate and its interaction with negative equity. It is shown that zip-level foreclosure rate itself is significant but negatively correlated with the probability of delinquency (the negative β_2). This tends to support the game-theoretic view that borrowers take nearby foreclosures as an indication of the chance of being foreclosed should she chooses to default, and thus nearby foreclosures lower the neighboring borrower's likelihood of becoming delinquent on her loan. But as just discussed, this variable can also be measuring some unobservable neighborhood characteristics that are orthogonal to other measured changes in the neighborhood. Therefore, I do not want to over-interpret this result. The clearer inference should be from the interaction term. The coefficient of

$ForclRate_{j,t} \cdot NegEq_{i,t}$ is positive and significant (the positive β_3) meaning that the higher the foreclosure rate is in neighborhood, the more sensitive borrowers are to negative equity in their delinquency choice. This positive net impact of neighborhood foreclosure concentration on delinquency suggests that the foreclosure contagion effect likely outweighs the information effect.

To account for possible changes in borrower's sensitivity to negative equity due to other reasons such as the overall market sentiment, we include the interaction of current year dummies with negative equity in Model 2. It is observed that there is no material change to the results I just discussed. β_3 is still positive and highly significant.

Table 4.5 results of our models where we use a dummy variable to indicate where a specific zip code during a specific quarter has high foreclosure rate comparing to other zip-quarters. Here high foreclosure rate means that it is in the 90th percentile of all zip-quarters. Other than this change, the model specification is exactly the same here in Table 4.5 as in Table 4.4. Results are consistent with those in Table 4.4. β_3 is still positive and highly significant, suggesting strong positive net impact of neighborhood foreclosure concentration on borrowers' propensity to exercise default option.

Table 4.4 MLE Estimates of the Cox Proportional Hazard Model

Covariate	Estimate	Estimate
Negative equity	0.655***	2.61***
	(0.144)	(0.177)
Negative equity square	0.003***	0.011***
	(0.000)	(0.002)
Negative equity *Zip-level Foreclosure unit	0.192***	0.169***
	(0.04)	(0.042)
Zip-level Foreclosure unit	-0.069**	-0.052*
	(0.028)	(0.029)
Negative equity *Low/no doc indicator	0.088*	0.048
	(0.046)	(0.042)
Low/no doc indicator	0.175***	0.179***
	(0.029)	(0.029)
Negative equity *Owner-occupied property indicator	-0.259*	-0.226*
	(0.141)	(0.135)
Owner-occupied property indicator	0.065	0.059
	(0.082)	(0.081)
Negative equity *FICOSCORE	0.176***	0.136***
	(0.017)	(0.016)
FICOSCORE	-0.132***	-0.116***
	(0.012)	(0.012)
FICOSCORE*FICOSCORE	0.041***	0.041***
	(0.004)	(0.004)
Log balance	0.117***	0.09***
	(0.016)	(0.017)
LTV at origination >=80%	0.074**	0.078**
	(0.035)	(0.035)
Call option in the money but covered by prepayment	0.066***	-0.015
	(0.011)	(0.013)
Call option in the money and out of prepayment penalty	0.006	0.006
	(0.007)	(0.007)
15-year FRM	0.067	0.049
	(0.062)	(0.062)
Planned-unit development	-0.127*	-0.121**
	(0.066)	(0.066)
Condominium	-0.025	-0.044
	(0.044)	(0.044)
Rate/term refi	-0.471***	-0.473***
	(0.057)	(0.057)
Cash out refi	0.113**	0.056
	(0.05)	(0.05)
With prepayment penalty clause	0.203	0.186
	(0.154)	(0.154)

Table 4.4 MLE Estimates of the Cox Proportional Hazard Model (Continued)

Covariate	Estimate	Estimate
Unknown prepayment penalty clause	0.121	0.128
	(0.156)	(0.157)
Change in MSA unemployment rate	0.322***	0.417***
	(0.02)	(0.024)
Payment-to-Income (PTI)	0.018**	0.016*
	(0.009)	(0.009)
Asian	-0.056	-0.035
	(0.051)	(0.051)
Black	0.064*	0.062*
	(0.037)	(0.037)
Other race	-0.025	-0.011
	(0.026)	(0.026)
Female	0.042*	0.037
	(0.024)	(0.024)
Time Fixed Effects	NO	Current year-fixed effect in negative equity beta
N	263,656	263,656
-2LogL	136,406	135,952
AIC	136,462	136,036

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA, during the period 1999-2013. Negative equity is calculated with the contemporaneous house value (based on MSA level HPI) and the market value of the mortgage loan outstanding, adjusted by MSA-level house price volatility. Zip-level Foreclosure unit is calculated as the permillage of the total number of foreclosures in the past two quarters (e.g. for 2009Q1 it is 2008Q4 and 2008Q3) in the total number of housing units in each zip code. Log balance refers to the log of the original loan amount. Call option is computed the difference between the par value of the mortgage and the present value of the remaining payments evaluated using the current market mortgage rate. Change in MSA unemployment rate refers to the difference between the unemployment rate at current time and at origination time. The other explanations of the variables are shown in Table 4.3. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Table 4.5 Hazard Model Estimates based on Alternative Neighborhood Foreclosure Concentration Measure

Covariate	**Estimate** **(S.E.)**	**Estimate** **(S.E.)**
Negative equity*High Foreclosure Intensity	0.633*** (0.124)	0.626*** (0.125)
High Foreclosure Intensity	-0.255*** (0.079)	-0.241*** (0.079)
Time Fixed Effects	NO	Current year-fixed effect in negative equity beta
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, Asian borrower, African American borrower, other non-white race borrower, female borrower.	
N	263,887	263,887
-2LogL	136,410	135,947
AIC	136,466	136,033

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA based on alternative neighborhood foreclosure concentration measure, during the period 1999-2013. High Foreclosure Intensity equals one if the zip-quarter ranks in the 90th percentile of all zip-quarters for its foreclosure intensity. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

4.4.2 Impact of foreclosure concentration by different regimes

It is reasonable to assume that the information effect of neighborhood foreclosure concentration to be stable over time. However, the contagion effect might vary in different regimes. Before 2007, the housing market was glorious. There were very few foreclosures and foreclosure was not a serious concern.

Therefore, the foreclosure contagion effect is expected to be minimal. These are exactly what the next a few tests show. In Table 4.6, results of a model are shown where the impact of neighborhood foreclosure concentration is allowed to vary in different regimes. The whole study period is divided into four regimes: pre-2007 is the period of housing boom; 2007 to 2009 is when the market experiences the first wave of the housing and mortgage market crisis; 2010-2011 is when we had the second wave of the crisis during which Los Angeles had a second downturn in the housing market after a short recovery in the second half of 2009; post 2012 is when the Los Angeles housing market had a real recovery. The *net* impact of neighborhood foreclosure concentration is indeed negative pre-2007, consistent with the notion that foreclosure contagion effect was small if not zero while the information effect was significant and negative. During 2007-2009, the *net* impact turned positive, likely due to the fact that foreclosure contagion became significant and prevalent. In 2010 and 2011, the *net* positive impact became even stronger compared to that during 2007-2009, possibly because of stronger contagion effect due to the desperation brought by the second wave of the crisis. Finally, post-2012 the net impact is not significant, likely due to a balance of the information effect and the foreclosure contagion effect.

Table 4.6 Hazard Model Estimates w.r.t. Different Housing Market Regimes

Covariate	Estimate (S.E.)
Negative equity *Zip-level Foreclosure unit * Pre 2007	-0.317*** (0.090)
Negative equity *Zip-level Foreclosure unit*Yr2007_2009	0.208*** (0.051)
Negative equity *Zip-level Foreclosure unit*Yr2009_2012	0.369*** (0.085)
Negative equity *Zip-level Foreclosure unit*Post 2012	0.062 (0.06)
Zip-level Foreclosure unit * Pre 2007	-0.20** (0.086)
Zip-level Foreclosure unit*Yr2007_2009	-0.07** (0.035)
Zip-level Foreclosure unit*Yr2009_2012	-0.278*** (0.057)
Zip-level Foreclosure unit*Post 2012	0.251*** (0.051)
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, Asian borrower, African American borrower, other non-white race borrower, female borrower.
N	263,656
-2LogL	136,282
AIC	136,350

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA with respect to different housing market regimes, during the period 1999-2013. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

4.4.3 Impact of foreclosure concentration by different borrower groups

The information effect and contagion effect could also vary with respect to different borrower groups. This research conducts such tests subsequently. In order to avoid the confounding effect of housing market regimes, the tests are conducted with the post 2007 subsample. The first test is whether Asian borrowers behave differently from the rest of the population. Table 4.7 shows the results. Interestingly, it is shown that the *net* impact of neighborhood foreclosure concentration for Asian borrowers is significantly different from non-Asian borrowers. Both its impact on delinquency probability and its impact on borrower's sensitivity to negative equity are stronger among Asian borrowers than among non-Asian borrowers. A possibly explanation is that due to cultural differences Asians are more susceptible to herd behavior[48]. Table 4.8 shows the comparison between African American borrowers and the rest of the population. There is almost no difference between African Americans and non-African Americans.

[48] For example, Chiang and Zheng (2010) find stronger evidence of herding in Asian stock market than in the US and Latin American markets.

Table 4.7 Asian Borrowers vs. Non-Asian Borrowers

Covariate	Estimate (S.E.)
Negative equity * Zip-level Foreclosure unit *Non-Asian borrower	0.198*** (0.051)
Zip-level Foreclosure unit * Non-Asian borrower	-0.078** (0.035)
Negative equity * Zip-level Foreclosure unit * Asian borrower	0.674*** (0.236)
Zip-level Foreclosure unit * Asian borrower	-0.465*** (0.172)
Non-Asian borrowers	--
Asian borrower	-0.092 (0.079)
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, African American borrower, other non-white race borrower, female borrower.
N	165,747
-2LogL	110,641
AIC	110,905

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA by comparing Asian and non-Asian borrowers, during the period 2007-2013. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Table 4.8 African American Borrowers vs. the Rest of the Population

Covariate	Estimate (S.E.)
Negative equity * Zip-level Foreclosure unit * Non-African American borrower	0.208*** (0.056)
Zip-level Foreclosure unit * Non-African American borrower	-0.086** (0.038)
Negative equity * Zip-level Foreclosure unit * African American borrower	0.202** (0.082)
Zip-level Foreclosure unit * African American borrower	-0.086 (0.064)
Non-African American borrower	--
African American borrower	0.079 (0.056)
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, Asian borrower, other non-white race borrower, female borrower.
N	165,747
-2LogL	110,647
AIC	110,707

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA by comparing African American and the rest of the population, during the period 2007-2013. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Next, female borrowers and male borrowers are compared. Interestingly, seen from Table 4.9, females have smaller β_3, suggesting that either the contagion effect is weaker or the information effect is stronger among female borrowers. A possibly explanation is that females have higher opportunity cost of homeownership and are more concerned with the negative consequences of foreclosure, which makes the information effect to be stronger and offsets more of the contagion effect.

Table 4.9 Female vs. Male Borrowers

Covariate	**Estimate** **(S.E.)**
Negative equity * Zip-level Foreclosure unit * male borrower	0.220*** (0.058)
Zip-level Foreclosure unit * male borrower	-0.092** (0.04)
Negative equity * Zip-level Foreclosure unit * female borrower	0.177** (0.07)
Zip-level Foreclosure unit * female borrower	-0.072 (0.049)
Male borrower	--
Female borrower	0.054 (0.036)
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, Asian borrower, African American borrower, other non-white race borrower.
N	165,747
-2LogL	110,646
AIC	110,706

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA by comparing female and male borrowers during the period 2007-2013. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

4.4.4 Impact of foreclosure concentration by different neighborhoods

Whether the foreclosure concentration effects vary in different neighborhoods is further analyzed. First neighborhoods are classified by average FICO score. For each zip code, the average FICO score of fixed-rate subprime mortgage loans originated during our study period and rank order all zip codes in the Los Angeles MSA are calculated. Then a dummy variable is used to indicate whether a neighborhood is in the upper or lower quartile in average FICO score. Finally, these dummy variables are interacted with our focus variables. Table 4.10 shows the model results. Interestingly, the results show that there is a U-shape in the relation between neighborhood average FICO score and the impact of foreclosure concentration on borrowers' sensitivity to negative equity. β_3 is significantly higher in very high and very low average FICO neighborhoods, while the middle tier FICO neighborhoods see decreased borrower sensitivity. Notice that it is already observed that borrowers with higher FICO score are more sensitive to negative equity (the positive coefficient of the interaction term between negative equity and FICO score), which is suggestive that borrowers with higher FICO score are more financially sophisticated and more responsive to financial opportunities. The finding that the neighborhood foreclosure concentration impact is more profound among high FICO neighborhood is consistent with such a financial sophistication explanation. In a separate test (shown in Table 4.11), neighborhoods are classified based on average income and find that lower-income neighborhoods see stronger relation between foreclosure concentration and borrower sensitivity to negative equity, while there is no significant difference between moderate-income neighborhoods and high-income neighborhoods in terms of the impact of foreclosure concentration on borrowers' sensitivity to negative equity.

Lastly, the analysis is generalized to the whole state of California. Results in Table 4.12 show that the impact of neighborhood foreclosure concentration in California is very similar to what we find in Los Angeles MSA.

A number of robustness tests are conducted including the use of different house price index to construct the negative equity measure as well as alternative foreclosure rate measure (e.g., per capital vs. per housing unit foreclosure rate). Results are robust.

Table 4.10 High FICO vs. Low FICO Neighborhoods

Covariate	Estimate (S.E.)
Negative equity * Zip-level Foreclosure unit * Lower_FICO	0.251*** (0.08)
Zip-level Foreclosure unit * Lower_FICO	-0.118** (0.057)
Negative equity * Zip-level Foreclosure unit * Middle_FICO	0.145** (0.06)
Zip-level Foreclosure unit * Middle_FICO	-0.021 (0.042)
Negative equity * Zip-level Foreclosure unit * Upper_FICO	0.315*** (0.095)
Zip-level Foreclosure unit * Upper_FICO	-0.193*** (0.067)
Lower_FICO	--
Middle_FICO	-0.067 (0.045)
Upper_FICO	0.037 (0.051)
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, Asian borrower, African American borrower, other non-white race borrower.
N	165,747
-2LogL	110,637
AIC	110,705

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA by comparing the lower, middle and upper quartiles of FICO SCORE at zipcode level during the period 2007-2013. Lower_FICO (Upper_FICO) equals one if the zip-quarter ranks in the 10th (90th) percentile of all zip-quarters for its FICO SCORE. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Table 4.11 High Income vs. Low Income Neighborhoods

Covariate	Estimate (S.E.)
Negative equity * Zip-level Foreclosure unit * Lower_Income	0.269** (0.132)
Zip-level Foreclosure unit * Lower_ Income	-0.206** (0.093)
Negative equity * Zip-level Foreclosure unit * Middle_ Income	0.188*** (0.059)
Zip-level Foreclosure unit * Middle_ Income	-0.06 (0.041)
Negative equity * Zip-level Foreclosure unit * Upper_ Income	0.203*** (0.07)
Zip-level Foreclosure unit * Upper_ Income	-0.082 (0.05)
Lower_ Income	0.137** (0.06)
Middle_ Income	0.03 (0.039)
Upper_ Income	--
Control variables	Negative equity, negative equity square, negative equity * low/no doc indicator, low/no doc indicator, negative equity * owner-occupied property indicator, owner-occupied property indicator, negative equity * FICO, FICO, FICO square, log loan balance, original LTV greater than 80%, call option value, 15-year FRM indicator, planned unit development indicator, condominium indicator, rate/term refinance indicator, cash-out refinance indicator, prepayment penalty indicator, prepayment penalty unknown indicator, change in MSA unemployment rate from origination to current, payment-to-income ratio, Asian borrower, African American borrower, other non-white race borrower.
N	165,747
-2LogL	110,639
AIC	110,707

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the Los Angeles MSA by comparing the lower, middle and upper quartiles of borrower median income at zipcode level during the period 2007-2013. Lower_Income (Upper_Income) equals one if the zip-quarter ranks in the 10th (90th) percentile of all zip-quarters for its median income. The other explanations of the variables are shown in Table 4.3 and Table 4.4. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

Table 4.12 Hazard Model Results for All Subprime Loans in the State of California

Covariate	Estimate	Estimate
Negative equity	0.442***	2.383***
	(0.072)	(0.095)
Negative equity square	0.004***	0.005***
	(0.000)	(0.001)
Negative equity *Zip-level Foreclosure unit	0.154***	0.113***
	(0.022)	(0.024)
Zip-level Foreclosure unit	-0.07***	-0.051***
	(0.013)	(0.013)
Negative equity *Low/no doc indicator	0.115***	0.033
	(0.032)	(0.029)
Low/no doc indicator	0.169***	0.191***
	(0.018)	(0.017)
Negative equity *Owner-occupied property indicator	0.001	-0.011
	(0.071)	(0.062)
Owner-occupied property indicator	-0.089**	-0.091**
	(0.041)	(0.04)
Negative equity *FICOSCORE	0.172***	0.122***
	(0.011)	(0.01)
FICOSCORE	-0.102***	-0.088***
	(0.006)	(0.006)
FICOSCORE*FICOSCORE	0.047***	0.05***
	(0.002)	(0.002)
Log balance	0.132***	0.128***
	(0.008)	(0.008)
LTV at origination >=80%	0.106***	0.108***
	(0.017)	(0.017)
Call option in the money but covered by prepayment	0.066***	0.025***
	(0.006)	(0.007)
Call option in the money and out of prepayment penalty	-0.004	-0.003
	(0.005)	(0.005)
15-year FRM	-0.093***	-0.098***
	(0.035)	(0.035)
Planned-unit development	-0.103***	-0.097***
	(0.035)	(0.035)
Condominium	0.019	0.031
	(0.031)	(0.031)
Rate/term refi	-0.381***	-0.378***
	(0.03)	(0.03)
Cash out refi	0.172***	0.134***
	(0.026)	(0.026)
With prepayment penalty clause	0.058	0.01
	(0.082)	(0.082)

Table 4.12 Hazard Model Results for All Subprime Loans in the State of California (Continued)

Covariate	Estimate	Estimate
Unknown prepayment penalty clause	-0.04	-0.065
	(0.084)	(0.084)
Change in MSA unemployment rate	0.31***	0.373***
	(0.011)	(0.012)
Payment-to-Income (PTI)	0.022***	0.022***
	(0.004)	(0.004)
Asian	-0.093***	-0.074**
	(0.032)	(0.032)
Black	0.036	0.043*
	(0.024)	(0.024)
Other race	-0.018	-0.011
	(0.014)	(0.014)
Female	0.034**	0.029**
	(0.014)	(0.014)
Time Fixed Effects	NO	Current year-fixed effect in negative equity
N	748,241	748,241
-2LogL	489,080	487,835
AIC	489,136	487,919

Note:

This table presents the Cox proportional hazard model result for the refined Subprime sample in the State of California, during the period 1999-2013. Negative equity is calculated with the contemporaneous house value (based on MSA level HPI) and the market value of the mortgage loan outstanding, adjusted by MSA-level house price volatility. Zip-level Foreclosure unit is calculated as the permillage of the total number of foreclosures in the past two quarters (e.g. for 2009Q1 it is 2008Q4 and 2008Q3) in the total number of housing units in each zip code. Log balance refers to the log of the original loan amount. Call option is computed the difference between the par value of the mortgage and the present value of the remaining payments evaluated using the current market mortgage rate. Change in MSA unemployment rate refers to the difference between the unemployment rate at current time and at origination time. The other explanations of the variables are shown in Table 4.3. Parameter estimates are reported standard errors are included in the parentheses. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.

4.5 Summary

Existing research has found foreclosure to be contagious in that foreclosure reduces the price of nearby non-distressed sales. This paper finds another type of foreclosure contagion – foreclosures can induce nearby mortgage borrowers to exercise their default option more ruthlessly. This type of foreclosure contagion is especially prominent during a downturn of the housing market. Therefore, during the mortgage market crisis, there are a large number of mortgage loans become delinquent, many of which subsequently were foreclosed. Those foreclosures were definitely bad results for the borrowers, the lenders and the investors. But the damage was not limited to the borrowers and lenders who are directly involved in the default process. Those foreclosures generate externalities to the neighborhood – they induce more borrowers in the surrounding area to enter into default. This circular reaction can go on and on and lead to foreclosure cascades. Therefore, it is important for the government and lenders to take timely actions to stop or reduce foreclosures and thus to break the loop of such a crisis.

Certainly, the impact of neighborhood foreclosure concentration on borrower default behavior is not limited to the contagion effect. It is actually shown that sometimes the impact can be on the opposite direction – foreclosures can discourage borrower's delinquency if borrowers take foreclosures as a signal of how lenders will deal with delinquencies. This information effect can dominate the contagion effect during the market boom. From this perspective, borrowers are strategic in their default decisions. Credit risk modelers thus should take this game feature of mortgage default into consideration to achieve better understanding and estimation of mortgage default risk.

Future research should try to establish the exact mechanism of the foreclosure contagion discovered in this paper, and assess the relative roles of observational learning, herding and other channels in generating such foreclosure contagion.

Chapter 5 Probabilistic Data Linkage in Real Estate Studies: Applications of Propensity Score Matching and Hard Matching with Machine Learning Techniques

5.1 Introduction

Since the recent few decades, data linkage, which is matching or integrating different datasets to identify the records of the same entity, is frequently used and has become an increasingly popular method in many fields, including statistics, economics, medicine (or public health), political science, sociology and even law, for the purpose of extending the amount of information available for the same entity and thus making decisions and taking actions (Gu et al., 2003).

The increasing popularity of data linkage arises from the fact that in most cases, the need for information often requires the analysis based on a large number of variables, which usually cannot be fulfilled by a single dataset. Data linkage can help exploit the information already available in different data sources, i.e. to carry out a statistical integration of information already collected. In addition, the advances in data processing techniques make data linkage technically and economically feasible to carry out the huge amount of operational work in integrating records from multiple datasets, thus to produce an enhanced dataset for research purposes.

However, there is no consensus on how exactly data linkage ought to be done, how to measure the success of the linking procedure and whether linking estimators are sufficiently robust to misspecification so as to be useful in practice (Heckman et al., 1998). Linking multiple datasets depends on the situations of those datasets and the requirements for information. Data linkage methods generally employ both deterministic and probabilistic linking algorithms (Jaro, 1995; Silveira and Artmann, 2009). If a unique identifier or key of the entity of interest is available in the record fields of all data sources to be linked, which is perfect but not commonly used in the current climate of data sources, the deterministic data linkage is used straightforwardly, by simply matching based on the identifier. If the unique key is not available, which is more complicated but more frequently encountered, the linkage becomes fuzzier since not all units can be unambiguously identified (Steiner and Cook, 2013).

One direct way for this fuzzy linkage, or probabilistic data linkage, is statistical hard matching, which identifies common covariates among different datasets and links these data using the common covariates. The other way is propensity score matching (PSM), which refers to the pairing of treatment and control units with similar values on the propensity score, and possibly other covariates, and the discarding of all unmatched units (Rubin, 2001). Both hard matching and PSM are commonly used in real estate studies, but neither method is perfect, regarding their advantages and disadvantages. In general, PSM is more suitable when dealing with a large number of covariates, whereas hard matching is more appropriate when dealing with a small number of covariates. However, both methods are limited in that they control for observed covariates but do not account for bias resulting from the unobserved covariates that may affect whether the entity from different data is the same or not and thus result in selection bias of linked records. This bias may in turn affect the results of analysis based on the biased data linkage. Also, PSM and hard matching both produce similar results when matching on a smaller number of covariates.

Mortgage data, the main data used in the whole thesis, have several characteristics that require an effective linking method. First, mortgage data usually track individual loans from certain lenders. Therefore, although there might be no unique identifiers among distinct mortgage data, loans from different data but the same group of lenders can be linked, if there are certain common attributes among these data. Second, the mortgage data provide distinct information at certain periods, i.e., loan application, loan origination, and loan termination. Thus, searching a way to link these data can help provide an overall idea about individual loan performance, probably through application to origination to termination. Third, because of data constraints and different collection methods, the common variables among the mortgage data might be just a few, thus might not capture the distinct characteristics of the loan records. The fact that the frequently used linking approaches mostly depend on common variables make it uncertain whether the linking is effective.

In this context, a comparative analysis of different methods of matching on multiple mortgage datasets is made. The motivation is to understand the limitations and potential of different approaches and in particular the ones based on machine learning techniques (Michie et al. 1994, Mitchell 1997). This analysis is also motivated during working on the fourth chapter, when I try to link BBX and HMDA data to analyze the foreclosure concentration impacts on borrower default behavior. This study is done by a systematic study, comparison and combination with traditional statistical matching techniques. A multi-strategy approach is used where several algorithms are applied to the same data and their results compared to find the best model. This is justified by the fact that it is very hard to select an optimal model a priory without knowing the actual complexity of a particular problem or dataset. This study provides important insights into the nature of the problem and allows us to address some fundamental questions such as: By comparing various linking approaches, what is the appropriate one to link multiple datasets in real estate studies, especially mortgage studies, when there are no unique identifiers? Among the linkage approaches, how can we minimize the selection bias and identification errors? Past studies comparing different approaches to the matching problem have been sometimes rightly criticized for using only one technique, for not being done in a systematic way or for consisting of mainly anecdotal results. In this study, I try to overcome this problem by systematically analyzing a variety of methods on linking the same groups of datasets, including statistical hard matching, statistical hard matching with machine learning techniques, i.e. Naïve Bayes classifier and Decision Tree Classifier, and propensity score matching. Many of these algorithms and methods were originally used by statisticians, computer or physical scientists, but their applications nowadays have been used in many economics and finance applications. There are several other advantages in comparing different methods in the same study: the pre-processing of the data is more homogeneous and the results can be compared in a more direct manner.

In this study, the BBX into HMDA data are linked with the common covariates among these two datasets, using the three approaches: pure statistical hard matching as in the literature, statistical hard matching combined with machine learning techniques, and propensity score matching. Firstly, with the statistical hard matching, I use the SAS program to link the selected common attributes and link BBX and HMDA data, se well as checking for and eliminating observations with duplicate linking records. Secondly, combining with the statistical hard matching, in dealing with the duplicate linking records, classification algorithms such as Naive Bayes and Decision Tree are applied to learn the intrinsic correlation of the key variables, including but not limited to selected common attributes, and produce the learned models for identifying the true matches. Thirdly, with propensity score matching, the BBX and HMDA records with the exactly same propensity scores or with the same three digits after the decimal point of propensity scores are regarded as match; the SAS program is applied to check for and eliminate observations with duplicate BBX id when there are multiple matches. Three groups of linked results are generated and compared accordingly. In the next step, to make sure the absence of sample selection problem, several checks for the matched samples are conducted, including the approaches frequently used by other economic studies such as distributions of the key variables, and summary statistics comparison. The bootstrapping approach is also used to estimate the accuracy rate of predicting the outcome of the loans for each method.

As a result, under the pure statistical hard matching and the matching with machine learning, the BBX-HMDA one-to-one exact matched sample forms the basis for the sample of about 2.5 million loans used in the analysis, which accounts for around 20% of the original BBX dataset. With the help of the SAS program to check for and eliminate observations with duplicate BBX id in multiple matches under pure statistical hard matching, there remain 2.6 million “actual” matches with unique BBX id from original multiple matches, which is 21% of the original BBX data. By applying machine learning techniques, it is shown that the model from Decision Tree Classifier obtains higher estimation score (81%) than Naïve Bayes Classifier (72%), inferring that Decision Tree Classifier better fits the data situation, and the high probability confirms that the one-to-one matched sample can be considered as true match exempted from misclassification problem. The trained model further classifies around 18% matches with unique BBX id from the original multiple matches. With regards to propensity score matching, it obtains only 26 thousand matches (less than 1% of the original BBX data) with exactly same propensity scores, but generates another 4.6 million matches (round 36% of the original BBX data) when the match is based on approximate propensity scores. Therefore, by comparing the number of linkages under the three approaches, it is observed that using statistical hard matching (Group 1) obtains slightly more linking records than using statistical hard matching with machine learning (Group 2), followed by propensity score matching (Group 3).

Next, in assessing the representativeness of the linked samples to the entire population of loans, several representativeness analyses are conducted on the linked groups, including examining the distributions of the key variables (by looking at kernel density distributions for continuous attributes and frequency plots for categorical attributes), comparing the summary statistics of the matched and original samples, and conducting the bootstrapping analysis on the outcome of default based on the key variables from both datasets. Results in general show that statistical hard matching with machine learning approach did a better job in dealing with selection bias and misclassification relative to the traditional approaches used in social science, such as pure statistical hard matching and propensity score matching. However, it is also shown that the performance of statistical hard matching, while not the best, is generally acceptable when there are no alternatives. Propensity score matching, although well packaged in various programs, should be used more carefully.

In conclusion, this study is expected to apply the commonly accepted techniques in statistics and computer science to linking records from multiple datasets in real estate field. This research is potentially useful in filling in additional or missing information, by adding in extra attributes. With more complete information on population units more complex research questions can be further addressed. Linking multiple datasets might be a way of checking accuracy and reliability of survey or administrative data or vice versa; one can assess whether the sample survey data are producing reliable inferences using some population administrative datasets to assess the representativeness of the sample data. Last, it can help enhance data quality, by providing more information for people to understand the non-response or non-report side of the current data. This study also extends the literature by comparatively analyzing different linking methods on multiple mortgage datasets to help better understand the advantages and potential limits of each method, as well as trying to overcome the selection bias and misclassification issues. These attempts help provide a creative way to other authors who also need to link multiple data sources with no unique identifiers and conduct deep analysis based on a representative matching dataset.

The rest of the chapter is organized as follows: the next section proposes data sources for studies and data preparation procedure. Section 5.3 develops the three methodologies used this study. Empirical results are discussed in the fourth section, while representative analysis is presented in the Section 5.5. Finally, concluding remarks are drawn in the last section.

5.2 Data Description

Two data sources are matched: loan-level data furnished by BlackBox Analytics (BBX) and the database of home loan applications and originations collected under the Home Mortgage Disclosure Act (HMDA).[49] The BBX database contains information on home location, mortgage amount, loan terms, and loan purpose. The HMDA data requires lenders to report data on borrower demographics, income, and geographic location for almost all loan applications in the United States. Therefore, the basic assumption is that most BBX mortgages should be contained in the HMDA database. The matched loans are identified using the common data fields across the databases. This analysis is limited to loans originated between 2001 and 2010.

5.2.1 BlackBox Logic (BBX)

Our estimates first rely on micro loan-level data set—BBX, which aggregates data from mortgage servicing companies that participate in their servicing agreement. The most recent BBX data cover about 22 million mortgages throughout the United States. The BBX dataset provides extensive information about the loan, property, and borrower characteristics at the time of origination as well as dynamically updated monthly data on loan performance subsequent to origination. Property-related variables are property appraisal value, geographic location (at zipcode level), and property type (single-family residence, condo, or other type of property). Loan characteristics available to us are loan amount at origination, interest rate type (whether the mortgage is fixed-rate or an adjustable-rate product), term to maturity, lien position, loan purpose (whether the loan was intended for home purchase or refinancing), and lender-defined subprime flag, the

[49] The Home Mortgage Disclosure Act (HMDA) was enacted in 1975, and implemented by the Federal Reserve Board. It requires that lending institutions report virtually all mortgage application and loan data. See http://www.ffiec.gov/hmda/ for details.

documentation type of the mortgage (full, low or no documentation), whether the loan was originated for an investor as well as coupon rate on the mortgage. Credit-risk-related variables include FICO credit score and loan-to-value (LTV) ratio of the borrower at origination. This study uses the loan-level data from BBX for loans originated in 2001-10.

Although BBX has substantial loan-level attributes, there are limited demographic or borrower-related information, which are significant in understanding the borrower and lender characteristics at the earlier stage of loan application and at the later stage of loan outcomes (if the loan application is accepted). Therefore, the data with Home Mortgage Disclosure Act (HMDA) files are considered due to its valuable information on the income of all loan applicants, in addition to race and various loan characteristics.

5.2.2 Home Mortgage Disclosure Act (HMDA)

The second dataset used is the Home Mortgage Disclosure Act (HMDA), which provides information on prime market share at MSA level. Under the HMDA data, most originators report basic attributes of the mortgage applications that they receive in metropolitan statistical areas to the Federal Financial Institutions Examination Council. This data is considered the most comprehensive source of mortgage data, covering around 80 percent of all home loans nationwide and a higher share of loans originated in metropolitan statistical areas (Avery et al., 2007).

HMDA provides abundant information about borrower and lender characteristics at the stage of loan application. Borrower characteristics incorporates applicant race, applicant sex, annual income and borrower-reported homeownership status (owner-occupied or investment). In terms of lender knowledge, HMDA indicates the application status for each applicant (denied or approved/originated). We can also gain the information of lender differences; for example, the number of loans originated by a lender in a given year (Agarwal et al. 2012). HMDA data provides a nearly complete universe of 122 million U.S. mortgage applications over the period 2001–10. Property-related variables available are geographic location (census tract level identification), and property type (one-to-four-family or manufactured housing or multifamily). HMDA also includes some but not much loan information, such as loan amount (in thousands), loan purpose (home purchase or refinancing or home improvement) and (in the case of originated loans) whether the loan was sold to the secondary market within the year.

In summary, both datasets are nationally representative and have been used by many researchers as major sources of information on mortgage analysis or potential behavior of borrowers, but they seldom have been used together, due to the lack of common record identifiers. The gap and significance in the literature motivates me to combine these two data sets using various linking approaches.

Table 5.1 provides definitions of the variables that are used in this study, for both BBX data and HMDA data.

Table 5.1 Variable list and definitions

Variable	Variable explanations
Panel A Variables from BBX data	
Original loan amount (*1000)	The amount of principal on the closing date of the mortgage (in thousands).
FICO score	The FICO (formerly the Fair Isaac Corporation) borrower credit score at the time of loan closing.
Original term	The number of months between the first payment date and the date the principal is due from the borrower.
Issuance balance	The coupon rate charged to the borrower for the initial remittance period.
Original LTV	The ratio of the original loan amount to the property value at loan origination; LTV is short for loan-to-value ratio.
Combined LTV	The ratio of all loan amounts on the property at the time of origination to the property value at loan origination.
Original appraisal value (*1000)	The estimate of the property value at the time of loan origination, as supplied by the data provider (in thousands).
Current interest rate	The coupon rate charged to the borrower for the most recent remittance period.
D_Second lien	A dummy that is equivalent to 1 for second lien loan, which is subservient to the main or first mortgage on a piece of real estate property; 0 otherwise.
D_Subprime	A dummy that equals to 1 if it is a subprime loan, defined by a lender.
D_Heloc	A dummy that is equivalent to 1 if it is a loan in which the lender agrees to lend a maximum amount within an agreed term, where the collateral is the borrower's equity in his/her house (HELOC is short for home equity line of credit). Otherwise it takes 0.
D_Interest only loan	1 if it is a loan in which, for a set term, the borrower pays only the interest on the principal balance with the principal balance unchanged, 0 otherwise.
D_FRM	1 for fixed-rate mortgages, 0 otherwise.
D_Prepayment penalty	1 if a fee will be charged to the borrower if they elect to make unscheduled principal payments, 0 otherwise.
D_condo	1 if the property securing the mortgage is condominium, 0 otherwise.
D_Single family	1 if the property securing the mortgage is single family, 0 otherwise.
D_Option ARM	1 if it is an adjustable rate mortgage with added flexibility of making one of several possible payments on your mortgage every month, 0 otherwise.
D_Purchase loan	1 if the primary reason the mortgage was take out by the borrower is to purchase, 0 otherwise.
D_Refinance loan	1 if the primary reason the mortgage was take out by the borrower is to refinance, 0 otherwise.
D_Owner occupied loan	1 if the use of the property is owner occupied (primary residence), 0 otherwise.
D_Investment loan	1 if the use of the property is investment, 0 otherwise.
D_Full documentation	1 if the amount of property documentation provided by the borrower is full documentation, 0 otherwise.
D_Low/No documentation	1 if the amount of income documentation provided by the borrower is low or no documentation, 0 otherwise.

Table 5.1 Variable list and definitions for this study (Continued)

Variable	Variable explanations
Panel B Variables from HMDA data	
Application status	1 if the applicant got acceptance of the loan, 0 otherwise.
Applicant race	Indicating the race of the applicant: i.e., American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian and White.
Applicant Ethnicity	Indicating the ethnicity of the applicant: i.e., Hispanic or Latino, Not Hispanic or Latino
Applicant sex	Indicating the sex of the applicant: i.e., male and female.
Applicant annual income	Gross Annual income of the applicant, in thousands of dolloars.
Loan type	The type of loan, defined by the lender: i.e., Conventional (any loan other than FHA, VA, FSA, or RHS loans), FHA-insured (Federal Housing Administration), VA-guaranteed (Veterans Administration), and FSA/RHS (Farm Service Agency or Rural Housing Service).
Property type	The type of the property securing the mortgage: i.e., one to four-family, manufactured housing, multifamily
Loan purpose	The primary reason the mortgage was take out by the borrower: i.e., home purchase, home improvement, refinancing.
Owner-occupancy status	The use of the property: i.e., owner-occupied as a principal dwelling, not owner-occupied.
Lien status	The relative claim position on a given property being used as collateral for a loan: i.e., secured by a first lien, by a subordinate lien, not secured by a lien.

Notes:

1. This list of variables only includes those used in this study, not all variables.
2. The variables with "D_" represent dummies.

5.2.3 Data Preparation and Harmonization

Data linkage needs to happen on linking on key variables, which both datasets have in common. Since no formal identifiers such as unique serial numbers are observed between these two datasets, informal key attributes are used as an identifier for the purposes of linking. The common variables in both datasets have to be aligned to each other in terms of definitions and measurement, and their distributions should be made comparable so that at the very least the two datasets do not differ significantly by means of the common variables (Kum and Masterson, 2008). Since the datasets used are intended to be representative at the national level, a very close correspondence between the two files in terms of the common variables is reasonably expected. Exceptions to this rule are generally the result of non-exact correspondence between actual records the two datasets have and this inevitably introduces error into the matching procedure due to mismatched samples.

The selection of the specific common variables for matching should be made carefully to maximize the explanatory power. This is because the validity of matching relies heavily on the power of the common variables to act as good predictors that can be transformed into effective informal identifiers. The common attributes between the two datasets are lien status, property type, loan purpose, occupancy status, original loan amount, origination year, and zipcode.[50][51] Among these common attributes, property types in BBX and HMDA data are defined differently.[52] Lien status is also questionable as a linking attribute, due to its incompleteness in HMDA data. Therefore, for this study the common variables used throughout different methods are loan purpose, occupancy status, original loan amount, origination year and zipcode.

Several filtering criteria are applied before actual linking. First, only loan applications marked as originated in HMDA data are considered. Those loans originated by FNMA, GNMA, FHLMC and FAMC are removed. Those with loan

[50] With regard to HMDA data, the Loan Application Date and Loan Action Taken Date (as well as the Loan Number) are considered non-public fields and are not released in any of the public FFIEC data products. The individual raw HMDA loan data are only available on an annual basis.

[51] There are no zipcodes in original HMDA data, but it contains census tract information. Therefore, using the census tract it is easy to identify the zipcode for each record.

[52] For instance, condominium loans are an independent category in BBX, while included in 1-4 unit family loans in HMDA.

type of FSA (Farm Service Agency) or RHS (Rural Housing Service) are excluded as well.

Table 5.2 compares the frequency distributions of the common variables used in linking procedure, in BBX and HMDA data respectively. The results show that the distributions of the three variables, loan purpose, occupancy status and original loan amount, are similar between BBX and HMDA, which infers that the two datasets are comparable, with respect to the key common attributes.

Table 5.2 The Frequency Distributions of Key Common Variables used in Data Linkage: BBX vs. HMDA

Loan purpose	**BBX**	**HMDA**
Home purchase	41.69%	38.53%
Rate/term refinance or Cash-out refinance	47.63%	53.32%
Unknown and other	10.68%	8.15%
Occupancy status	**BBX**	**HMDA**
Owner-occupied	81.12%	89.58%
Investment property	9.19%	9.90%
Unknown and other	9.69%	0.52%
Original Loan Amount	**BBX**	**HMDA**
Less than $10K	28.59%	37.33%
$100K-$300K	43.52%	46.85%
$300K-$500K	17.82%	11.12%
Greater than $500K	10.06%	4.70%

Note:

This table shows the frequency distributions of three key variables, loan purpose, occupancy status and original loan amount, by comparing BBX and HMDA data. The percentage number is shown and compared.

5.3 Methodology

In this study, a variety of methods on linking the same groups of datasets are tested and compared, including statistical hard matching, statistical hard matching with machine learning techniques, i.e. Naïve Bayes Classifier and Decision Tree Classifier, and propensity score matching.

5.3.1 Statistical hard matching (Method 1)

The design of statistical hard matching (the "*Method 1*") in this study is similar with other hard matching studies in real estate field (Agarwal et al, 2012; Agarwal et al., 2012; Ferreira and Gyourko 2011; Ghent et al., 2011; Haughwout et al., 2009; Hernandez-Murillo and Sengupta 2012; Pace and Zhu, 2012; Reid and Laderman 2009; Voicu et al., 2011). In the first stage, the BBX loans are matched to HMDA loans with the same loan purpose, occupancy status of the borrowers, loan origination year, zipcodes, and original loan amount. Any BBX loan that has no corresponding HMDA loans using the criteria in Section 5.2.3 is a non-match. Any loan is a multiple match if it matches to multiple HMDA loans. Lastly, any BBX loan that matches to one and only one HMDA loan is a one-to-one match. All one-to-one matches with unique BBX and HMDA identifiers are kept, while all non-matches are excluded from the final matched sample. For multiple matches, the SAS program is applied to check for and eliminate observations with duplicate linking records. The linking results using *Method 1* is regarded as *Group 1* in the following section.

5.3.2 Machine learning techniques (Method 2)

The second approach is the statistical hard matching combined with machine learning techniques (the "*Method 2*"). Machine learning is a key technique that exploits the nature of the dataset, e.g., the underlying patterns and relationship of variables. Classification algorithms such as Naive Bayes and Decision Tree are commonly used in the field of computer science, medicine, statistics, etc. The classifiers first learn the underlying pattern (term as model) from a set of labeled data (term as training set), and then apply the model to the unseen data (termed as predicting set) and predict the label (matched or non-matched in this analysis) for this predicting set. To compare the performance of different classifiers, I also apply the models to a small set of labeled data (term as testing set) and report their accuracy in this set.[53]

5.3.2.1 Naïve Bayes Classifier

A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.[54] In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood.

The probability model for the classifier is a conditional model:

[53] In statistical modeling, a training set is used to fit a model that can be used to predict a "response value" from one or more "predictors" from which it can construct or discover a predictive relationship. A test set is a set of data that is independent on the training data but follows the same probability distribution as the training data. A predicting set is a set of data that is unknown about the classification. A test set is a predicting set in certain situation.

[54] For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. A naïve Bayes classifier considers all of these properties to contribute to the probability that this fruit is an apple.

$$p(C \mid F_1, F_2, ..., F_n) \qquad (5.1)$$

over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F_1 through F_n. Using Bayes' theorem, and under the "naïve" conditional independence assumptions that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$ given the category C, this can be written as[55]

$$p(C \mid F_1, F_2, ..., F_n) = \frac{p(C)p(F_1, F_2, ..., F_n \mid C)}{p(F_1, F_2, ..., F_n)} \qquad (5.2)$$

$$= \frac{p(C, F_1, F_2, ..., F_n)}{p(F_1, F_2, ..., F_n)}$$

$$= \frac{p(C)\prod_{i=1}^{n} p(F_i \mid C)}{p(F_1, F_2, ..., F_n)}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features F_i are given, so that the denominator is effectively constant. Therefore, the conditional distribution over the class variable C is

$$p(C \mid F_1, F_2, ..., F_n) = \frac{1}{Z} p(C)\prod_{i=1}^{n} p(F_i \mid C) \qquad (5.3)$$

[55] In plain English the above equation can be written as $posterior = \frac{prior \times likelihood}{evidence}$.

where Z is the scaling factor dependent only on $F_1, F_2, ..., F_n$, that is, a constant if the values of the feature variables are known. For further explanations please refer to the Appendix.

The matching problem in this analysis is identical to a binary classification issue, which the probabilities of classifying the record as “match” or “unmatch” based on probability distributions of the measured features:

$$\mathrm{P(match}|F_1, F_2, \ldots, F_m) = \frac{P(match)P(F_1,F_2, \ldots, F_m|match)}{P(F_1,F_2, \ldots, F_m)} \tag{5.4}$$

$$\mathrm{P(unmatch}|F_1, F_2, \ldots, F_m) = \frac{P(unmatch)P(F_1,F_2, \ldots, F_m|unmatch)}{P(F_1,F_2, \ldots, F_m)} \tag{5.5}$$

Based on the conditional independence assumption of naïve Bayes model, assume that each feature F_i is conditionally independent of every other feature F_j for j ≠ i . So the above equation becomes

$$\mathrm{P(match}|F_1, F_2, \ldots, F_m) = \frac{P(match)P(F_1|match)P(F_2|match)\ldots P(F_m|match)}{P(F_1,F_2, \ldots, F_m)} \tag{5.6}$$

$$\mathrm{P(unmatch}|F_1, F_2, \ldots, F_m) = \frac{P(unmatch)P(F_1|unmatch)P(F_2|unmatch)\ldots P(F_m|unmatch)}{P(F_1,F_2, \ldots, F_m)} \tag{5.7}$$

If there is identical number of observations in each class in the training set, we will have P(match) = P(unmatch) = 0.5 in the training sample. $P(F_1, F_2, \ldots, F_m)$ can be ignored since it is a positive constant so it is the same for any sample.

For each record in the test set, the probability of match and unmatch is compared based on equation (5.6) and (5.7): the larger probability predicts the matching status of the record. Then we compare the predicted matching status with the actual matching status and calculate the probability that these two are the same among the entire test set.

The major advantage of the naive Bayes classifier is its short computational time for training. In addition, since the model has the form of a product, it can be converted into a sum through the use of logarithms – with significant consequent computational advantages. However, due to its strict independence assumption, it actually contradicts the real world situation: the attributes may have intrinsic correlations with one another.

5.3.2.2 Decision Tree Classifier

Decision trees classifier, commonly accepted in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. This approach is the best known and most widely used learning methods in data mining applications. The goal is to create a model that predicts the value of a target variable based on several input variables. Samples are classified by sorting them down the tree from the root to the leaf node. The leaf node provides classification of the sample. Each non-leaf node in the tree specifies a test of one or more attributes of the sample. Each branch descending from a node corresponds to one of the possible values for this attributes. A sample is classified by starting at the root node of the tree, testing the attribute specified by this node, and moving down the tree branch corresponding to the value of the attribute in the given sample. This is repeated for the subtree rooted at the new node. The process continues until a leaf is encountered, at which the object is asserted to belong to the class named by the leaf (Quinlan, 1986). Decision trees can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree, called top-down induction of decision trees (Quinlan, 1993).[56]

Data comes in records of the form:

$$(x, Y) = (x_1, x_2, x_3, \ldots, x_{k,} Y) \tag{5.8}$$

The dependent variable, Y, is the "target variable" that we are trying to understand, classify or generalize. The vector x is composed of the input variables, $x_1, x_2, x_3, \ldots, x_k$, used to predict Y.

The determination of the node splitting is based on ID3 (Inductive Dichotomizer 3), which uses a single best attribute to test at each node of the tree for classifying the samples. A statistical property called information gain is used, to measure how well a given attribute separates the training examples according to their target classification.

In order to define information gain precisely, I begin by defining a measure

[56] A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion if completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data (Quinlan 1986).

commonly used in information theory, called entropy, that characterizes the (im)purity of an arbitrary collection of examples. If the target attribute can take on C different values, then the entropy of S relative to this C-wise classification is defined as

$$Entropy(S) = -\sum_{j=1}^{C} p_j \log_2 p_j \tag{5.9}$$

where $\sum_{j=1}^{C} p_j = 1$, and p_j is the proportion of S belonging to class j. Take a binary classification as an example: given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this classification is

$$Engropy(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2) \tag{5.10}$$

where p_1 is the proportion of positive examples in S and p_2 is the proportion of negative examples in S. As the data become purer and purer, the entropy value becomes smaller and smaller.[57]

The information gain, which measures the effectiveness of an attribute in classifying the training data, is simply the expected reduction in entropy caused by partitioning the examples according to this attribute:

[57] Notice that the entropy is 0 if all members of S belong to the same class. The entropy is 1 when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (5.11)$$

where A is an attribute, $Values(A)$ is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S \mid A(s) = v\}$). The first term in Equation (5.11) is the entropy of the original collection S, and the second term is the expected value of the entropy after S is partitioned using attribute A.

5.3.2.3 Comparison of the machine learning techniques

Bias measures the contribution to error of the central tendency of the classifier when trained on different data. Variance is a measure of the contribution to error of deviations from the central tendency. Learning algorithms with a high-bias profile usually generate simple, highly constrained models which are quite insensitive to data fluctuations, so that variance is low. Naive Bayes is considered to have high bias, because it assumes that the dataset under consideration can be summarized by a single probability distribution and that this model is sufficient to discriminate between classes. On the contrary, algorithms with a high-variance profile, e.g. decision tree classifier, can generate arbitrarily complex models which fit data variations more readily. As we are aware that data in different years have different nature, we need to train classification models for each year respectively, and use those models to obtain the classification for predicting datasets for each year respectively.

Under Method 2, the first step is the same as pure statistical hard matching. After obtaining all possible HMDA matches for each BBX loan, the BBX loans are then classified as non-matches, one-to-one matches, or multiple matches. Non-matches are excluded from the final matched sample as well.[58] Machine learning techniques are applied to check the accuracy of one-to-one matches, and select the possible true match from multiple matches.

To follow the convention in machine learning, I also split the matched data into training set, test set and predicting set in this study. The training set is a combination of 80% of the one-to-one matches and random selection of the same amount of the non-matches.[59] The distribution probabilities of the measured features to the classification as a match or non-match are calculated by the two techniques. These measured features include the common covariates used in the statistical hard matching as well as other attributes, i.e., property appraisal values, loan-to-value (LTV) ratio, borrower FICO score, subprime loan indicator, HELOC indicator, interest-only loan indicator, fixed-rate mortgage (FRM) indicator, prepayment penalty indicator, property type (condo, single family or multifamily), option ARM indicator, loan document type indicator (full documentation, low documentation or other), applicant income, race, sex, ethnicity, and loan type (conventional loan, FHA loan or other). The goal is to

[58] Although machine learning techniques can also be applied to non-matches to check the non-matches, it may result in additional selection bias and lots of uncertainty. Therefore, non-matches are excluded in this research to avoid unnecessary bias.

[59] The one-to-one matches used as part of the training set are randomly selected from the original one-to-one matches, using the SAS program, and the non-matches used are randomly selected from the original non-matches.

learn the intrinsic correlation of the key variables, including but not limited to selected common attributes, and produce the learned models.

The test set is the rest 20% of the one-to-one matches combined with the randomly selected 20% of the non-matches, but with the classification ("match" or "unmatch") hidden first. The learned models from the training data are applied to the test set to predict the classification. The accuracy rate of the models are calculated as the percentage of records having the same predicted classification as the hidden ones, based on all significant variables including but not limited to common variables. The higher the accuracy rate, the better the classifier fits to the matching sample, and thus the better the learned model fits to the actual data situation. This is also to confirm that the one-to-one matches do not suffer from misclassification problem in general. The predicting set in this research is the multiple matches, where the better learned model is applied to select the actual match. The linking results using *Method 2* is regarded as *Group 2* in the following section.

5.3.3 Propensity score matching (Method 3)

The third method is the propensity score matching (the "*Method 3*"), which is commonly applied in statistics, economics, medicine and other fields. The propensity score is a balancing score: conditioning on the true propensity score asymptotically balances the observed covariates. Propensity scores are used in observational studies to reduce selection bias by matching different groups based on these propensity score probabilities, rather than matching patients on the values of the individual covariates.

The estimated propensity score $p(x_i)$, for subject i, ($i = 1,...,N$) is the conditional probability of being assigned to a particular category given a vector of observed covariates x_i (Rosenbaum and Rubin, 1983):

$$p(x_i) = \Pr(z_i = 1 \mid x_i) \tag{5.12}$$

and

$$\Pr(z_1,...,z_n \mid x_1,...,x_n) = \underset{i=1}{\overset{N}{\tilde{\mathrm{O}}}} e(x_i)^{z_i}\{1 - e(x_i)\}^{1-z_i} \tag{5.13}$$

Where $z_i = 1$ for treatment, $z_i = 0$ for control, x_i is the vector of observed covariates for the i^{th} subject. Since the propensity score is a probability, it ranges in value from 0 to 1.

The vast majority of published propensity score analyses use logistic regression to estimate the scores. Logistic regression is attractive for probability prediction, since it is mathematically constrained to produce probabilities in the range (0, 1), and generally converges on parameter estimates relatively easily. Further, logistic regression is a familiar and reasonably well-understood tool of researchers in a variety of disciplines, and is easy to implement in most statistical packages (Westreich et al., 2009). In this study, logistic regression is applied to estimate the propensity scores.

The method of matching records in different data sources based on propensity scores demonstrated here is based on matching on an allowable absolute difference between exact propensity scores, or a “radius” around the score. This matching is done using a generalized SAS macro for propensity score matching that can match a “control group” to a “patient group” at an N:1 ratio, using an algorithm to maximize the number of propensity score matches (Fraeman, 2010). This optimization algorithm is based on retaining the matches for loan records with the fewest possible number of matches first.

Following Fraeman (2010), the procedure of applying propensity score matching in this study is basically two steps. In the first step, the BBX and HMDA data are aggregated into one data, with only BBX identifier, HMDA identifier and the selected common variables listed in Section 4.3.3; a new variable “source” is defined as 1 if the record is originally from BBX data, 0 if it is from HMDA data. The common variables, as the form of dummies, are included in the logistic regression to get the propensity scores, with dependent variable as “source”. Notice that since there are around 20,000 zipcodes among the data in each year, the constraints of the calculation matrix prevent the inclusion of all these zipcodes into the logistic regression at one time. In order to control for zipcode effects in such condition, records are separated into 100 groups, based on the zipcodes. [60] The BBX and HMDA records with the exactly same propensity scores are regarded as match; the SAS program is applied to check for and eliminate observations with duplicate BBX id when there are

[60] For example, records with zipcodes from 1,000 to 2,000 are classified as one group, et cetera.

multiple matches. The rest of the BBX and HMDA records are put into next round's matching. In the next step, the matching criterion is relaxed: other things being equal, the records with the same three digits after the decimal point of propensity scores are regarded as match. Records in BBX that cannot find corresponding matches in HMDA are excluded. For multiple matches, the SAS program is applied to check for and eliminate observations with duplicate BBX id. The linking results using *Method 3* is regarded as *Group 3* in the following section.

5.4 Empirical Results on Data Linkage

5.4.1 Basic linking results of Method 1 and Method 2

Both pure statistical hard matching and the matching with machine learning techniques return the same results of one-to-one matches, original multiple matches and non-matches. Under the matching algorithm in this study, the BBX-HMDA one-to-one exact matched sample forms the basis for the sample of about 2.5 million loans used in the analysis, which accounts for around 20% of the original BBX dataset. More than 2.6 million loans from BBX are multiply linked to HMDA data, similar with the number of one-to-one matches.[61] The remaining BBX loans are taken as non-matches which have no corresponding HMDA loans. The difference between Method 1 and Method 2 lies in the way to deal with multiple matches: how to select the match from multiple matches.

[61] The total multiple matches account for around 10 million linkage: for each record in BBX there are more than one record in HMDA that satisfy the matching criteria. On average, 4 records from HMDA are linked to every record in BBX.

In Method 1, which is the pure statistical hard matching, with the help of the SAS program to check for and eliminate observations with duplicate BBX id in multiple matches, there remain 2.6 million "actual" matches with unique BBX id from original multiple matches, which is 21% of the original BBX data. Therefore, combined with the one-to-one matched results, there are in total 5.1 million matches (40% of the original BBX), under the method of pure statistical hard matching.

In Method 2, by studying the potential influences of the variables on the classification from the training set and test set, the contributions (probabilities) of those variables to the final classification are calculated and a trained model is produced based on the probabilities. It is shown that comparing the two trained models obtained by the two techniques, the model from Decision Tree Classifier obtains the highest score (accuracy rate of around 81%), while that from Naïve Bayes Classifier gets around 72%. This result infers that Decision Tree classifier best fits the one-to-one matches' situation. This high probability also confirms that the one-to-one matched sample does not suffer from misclassification problem and can be considered as true match, while the randomly selected non-matches are the "actual" non-matches.

Based on the above results, the better learned model from Decision Tree Classifier is applied to the predicting set, which is the multiples matches, to find the "actual" matches. This trained model helps predict approximately 18% matches with unique BBX id from the original multiple matches. Therefore, together with one-to-one matches, there are around 4.8 million (38%) total matched sample. Detailed results are available upon request.

5.4.2 Data linkage result with Method 3

With regards to propensity score matching, in the first step with the exactly same propensity scores, it obtains only 26 thousand matches (less than 1% of the original BBX data). However, in the second step when the matching criterion is relaxed, another 4.6 million matches are generated, which are as round 36% of the original BBX data. In total, there are 4.6 million matches when applying propensity score matching, which account for more than 36% of the original sample. Results for propensity score matching are available upon request.

Table 5.3 shows the comparison of the linkage performance by looking at the number of linkages for the three groups: the original BBX sample, the linking records from statistical hard matching alone, statistical hard matching with machine learning, and propensity score matching. It is observed that among these three approaches, using statistical hard matching (Group 1) obtains slightly more linking records than using statistical hard matching with machine learning (Group 2), followed by propensity score matching (Group 3). This is due to the nature and linking mechanism of these three approaches: statistical matching relies on the program to keep only one record for each group of multiple matches, while the method with machine learning calculates the probability of correct linking based on common and uncommon variables. Thus those did not satisfy the probability requirement under machine learning have been excluded from the final linking sample. Finally, propensity score matching depends on the propensity scores for linking two records from BBX and HMDA, which may miss a group of linking records.

Table 5.3 Number of linkages for the three matched groups: 2001-2010

Year	Original BBX data	Group 1	Group 2	Group 3
2001	419,359	151,111	151,801	292,536
2002	679,599	268,906	248,785	376,318
2003	1,369,929	674,877	490,318	983,138
2004	2,133,744	953,077	925,079	810,999
2005	3,545,879	1,351,353	1,315,772	1,034,646
2006	3,458,331	1,274,137	1,245,401	745,992
2007	1,020,773	426,748	399,744	356,488
2008	3,601	1,713	1,707	1,380
2009	816	564	535	17
2010	477	268	269	41
Total	12,632,508	5,102,754	4,779,411	4,601,555
Percentage		40.39%	37.83%	36.43%

Note:

This table shows the number of linkages for the three groups, by year 2001-2010. The percentages of the linked sample to the original BBX data are calculated.

In the next section, the linking records from each method are examined carefully, using various representativeness analyses.

5.5 Representativeness Analysis

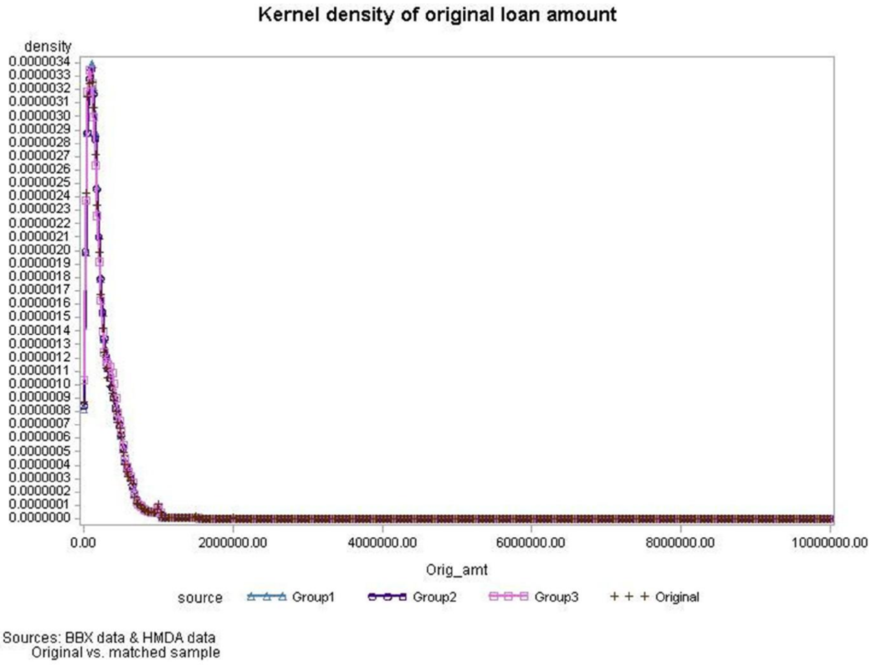
A closely related issue to the outcome of data linkage is whether the linked data A ∩ B is representative of a population in A ∪ B or subpopulation in data A or data B without bias in key parameters. An obvious issue is that any matching error in linking data A and data B can result in a ‘dirty’ sample set A ∩ B that might not be suitable for some (or most) analyses.

To assess the representativeness of the linked samples to the entire population of loans, several representativeness analyses are conducted, comparing the matched samples and original sample. The matched sample from pure statistical hard matching (*Method 1*) is named as *Group1*, statistical hard matching with machine learning techniques (*Method 2*) as *Group 2*, and the one drawn from propensity score matching (*Method 3*) as *Group3*.

5.5.1 Distributions of key variables

If the matched sample is representative of the whole population of loans, the distributions of the key variables from both datasets should be consistent between matched and original sample. Therefore, the key variables of the matched samples obtained from the three approaches are examined if they have similar distributions as those of original BBX and HMDA in general.

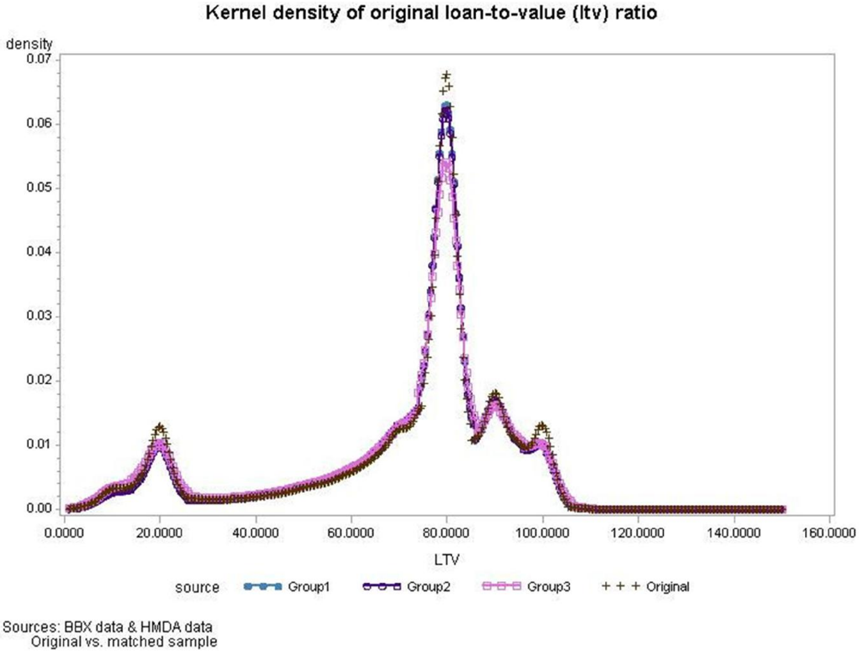
First the kernel density distributions of loan characteristics from BBX original dataset and Group 1 through Group 3 are checked, i.e. original loan amount, original loan-to-value (LTV) ratio, borrower FICO score (Figure 5.1). It is shown that the kernel density plots of these covariates reveal differences between the matched groups and the original BBX data. Shown in Figure 5.1.1 and Figure 5.1.3, with regard to original loan amount and FICO score, the kernel density distributions of the three groups are quite similar. However, concerning the distributions of original LTV ratio, Figure 5.1.2 shows that Group 1 and Group 2 are much closer to the distributions of the original BBX sample, compared with Group 3.


Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample

Figure 5.1 Kernel Density Plots of the BBX variables from the Original and Matched Samples: 2001-2010

Fig. 5.1.1 Kernel density plot of Original Loan Balance: original vs. matched groups


Kernel density of original loan-to-value (ltv) ratio

Sources: BBX data & HMDA data
Original vs. matched sample

Fig. 5.1.2 Kernel density plot of Original LTV: original vs. matched groups

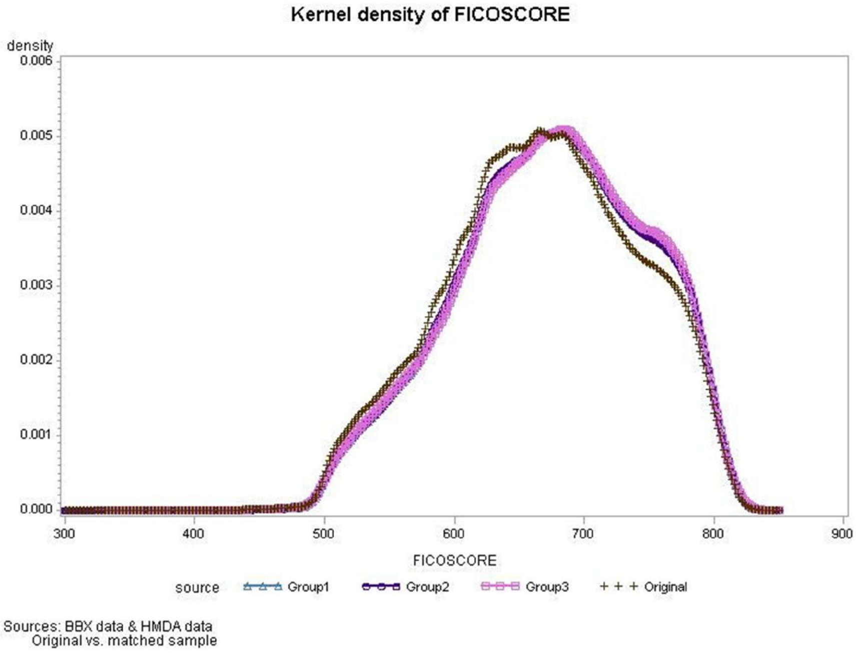


Fig. 5.1.3 Kernel density plot of FICO score: original vs. matched groups

Note:

This set of figures shows the kernel density plots of three key variables, original loan amount, original LTV ratio, and FICO score in BBX dataset. Four groups of data are compared: the original BBX dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Those datasets include loans originated over the period 2001-2010; only loans with original loan amount less than $10 million are included in the sample. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.

The distributions of the original loan amount, original LTV ratio and FICO score in each year are plotted as well, presented in the Appendix. Figure A1 shows the set of density plots for original loan amount for the four groups; before 2007, the distributions of the four groups are almost the same, while since 2008, the distributions of Group 3, which is under the propensity score matching, have gone away from those of the original BBX sample, while Group 1 and Group 2 still have the same trends. By looking at the distributions of the original LTV ratio for the four groups by year, Figure A2 reveals that except Year 2001, the distributions of Group 1 and Group 2 are more comparable to those of the original

BBX sample, with similar trends and smaller differences. This finding is consistent with that in Figure 5.1.2, which is the aggregate distribution of original LTV through 2001-2010. Presented in Figure A3, when looking into the distributions of FICO score by year, those of Group 3 are much more volatile than other groups and run away from the distributions of the original BBX sample, especially after Year 2007. These gradually increasing differences among the distribution plots confirm our thinking that propensity score matching tend to produce some bias when selecting the linking records.

With respect to the borrower information coming from HMDA dataset, the kernel density plots of borrower income, and the frequency distributions of applicant race, sex and ethnicity for original and matched samples are compared. Seen from Figure 5.2, the kernel density distributions of income for original HMDA sample, Group 1 and Group 2 have the same trend, while the distribution for Group 3 is quite distinct from all three groups. This result implicates that the matched results is not representative of the original HMDA data with respect to the borrower income information.

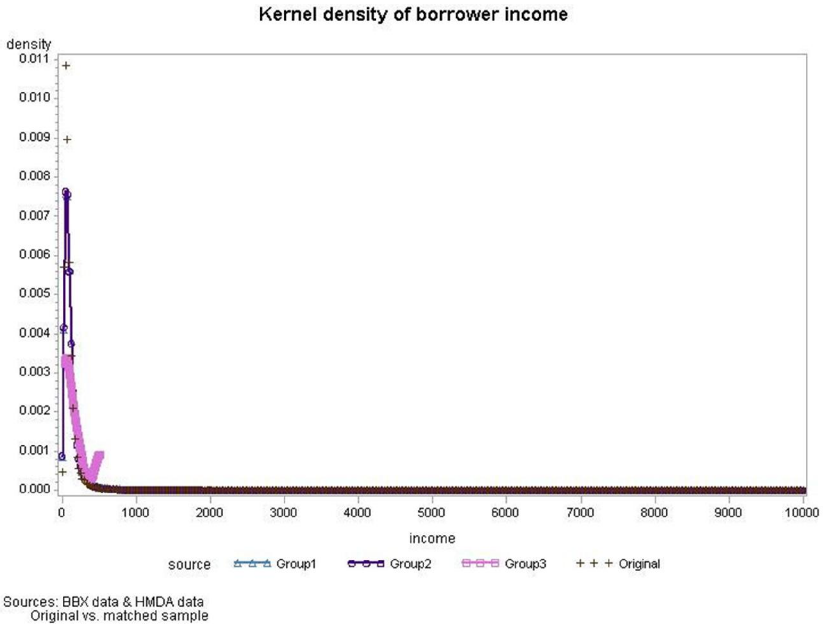


Figure 5.2 Kernel Density Plots of Borrower Income from the Original and Matched Samples: 2001-2010

Note:

This figure shows the kernel density plots of key variable, applicant annual income, in HMDA dataset. Four groups of data are compared: the original HMDA dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Those datasets include loans originated over the period 2001-2010; only loans with original loan amount less than $10 million are included in the sample. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.

For borrower characteristics, such as sex type, race type, and ethnicity type, the frequency distributions, comparing original HMDA dataset, Group 1, Group 2 and Group 3 are plotted. Shown in Figure 5.3, the frequency distributions of these attributes among the three groups are following the same trend, with slight differences in the percentages. The distributions of sex type (Figure 5.3.1) and race type in Group 2 (Figure 5.3.2), which is obtained from statistical hard matching with machine learning technique, are more similarly distributed with the original data, than Group 1 from pure statistical hard matching and Group 3 from propensity score matching. When comparing the ethnicity of the four groups, it is shown that the distribution of Group 1 is closer to that of the original HMDA sample than other groups.

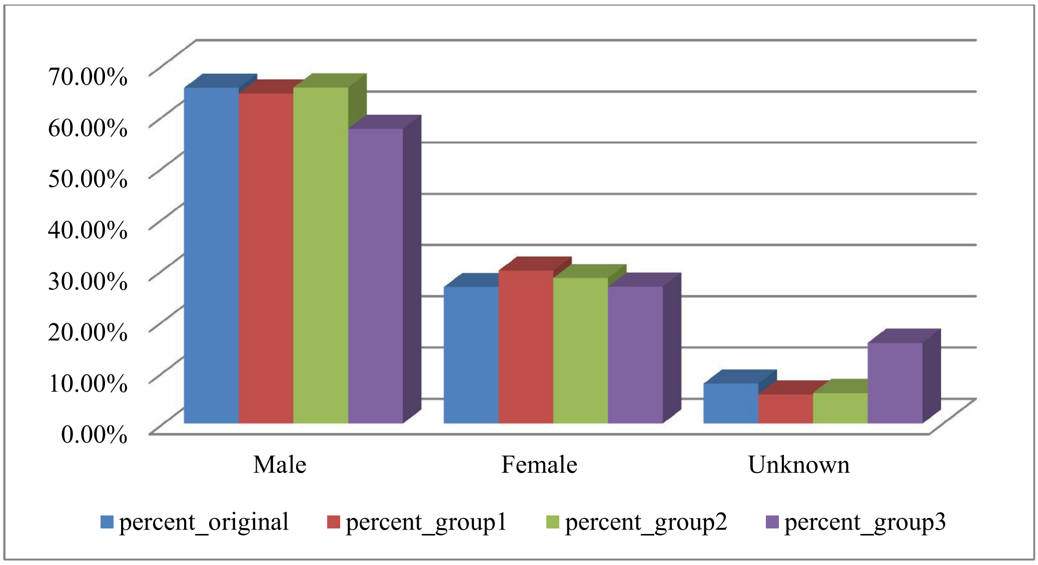


Figure 5.3 Frequency Distribution of Borrower Information from the Original and Matched Samples: 2001-2010

Figure 5.3.1 Frequency Distribution of Borrower Sex Type: original vs. matched groups

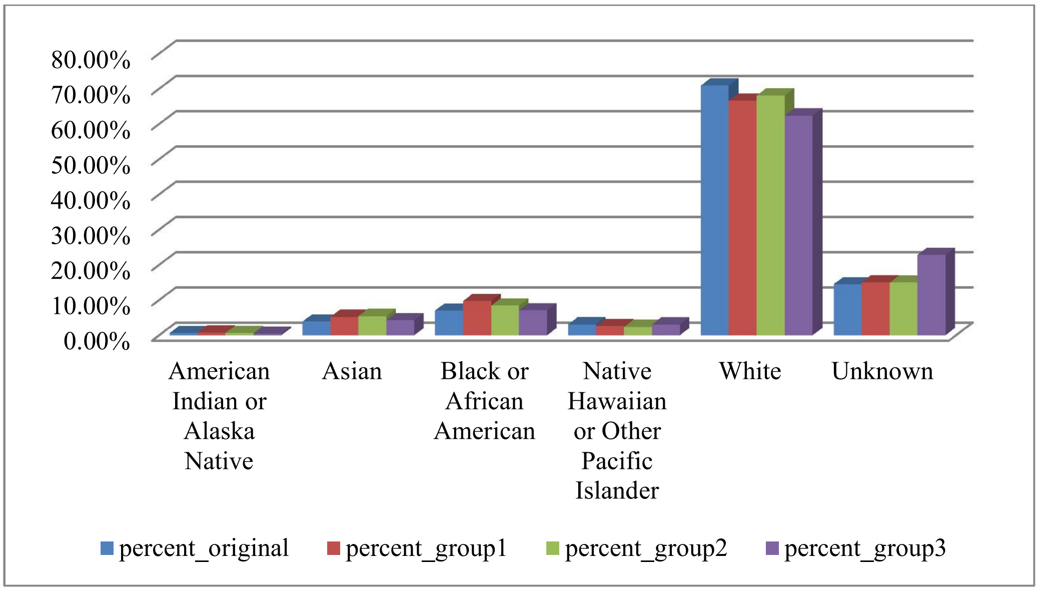


Figure 5.3.2 Frequency Distribution of Borrower Race Type: original vs. matched groups

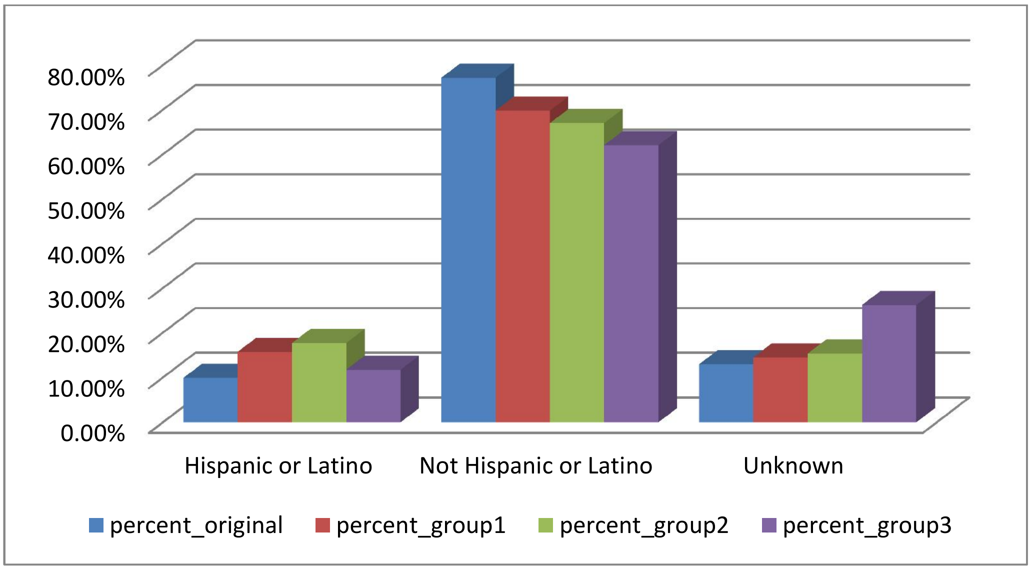


Figure 5.3.3 Frequency Distribution of Borrower Ethnicity Type: original vs. matched groups

Note:

This figure shows the frequency distribution of borrower variables (in percentages) of the loan application time. All the loans are originated over the period 2001-2010. Three groups of data are compared: the original HMDA dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Figure 5.3.1 shows the frequency distribution of borrower sex type; Figure 5.3.2 presents the distribution of borrower race type; Figure 5.3.3 displays the distribution of borrower ethnicity type. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.

In summary, these observations suggest that the sample under statistical hard matching with machine learning is generally more representative of the entire BBX sample and HMDA sample with respect to the key variables. Next the summary statistics comparisons are made for these groups, to present big pictures of the data.

5.5.2 Summary statistics comparisons

To check whether there is sample selection bias for the matched sample, another way frequently used in economic literature is to compare the summary statistics of the matched and original samples. Here the summary statistics of the original BBX dataset and the total matched samples, using three approaches, are compared.

Table 5.4 compares the descriptive statistics of the original BBX sample, total matched sample with pure statistical hard matching, with machine learning techniques, and under propensity score matching. In general, the gaps among the three groups are small, e.g. FICO score, original LTV, percentage of condominium loans, margin, and percentage of fixed-rate mortgage (FRM). Consistent with the assumptions, the gaps between original sample and the matched sample with statistical hard matching alone (Group 1) and with machine learning techniques (Group2) are smaller, relative to those between original sample and matched sample with random selection (Group3).

Table 5.4 Summary Statistics of BlackBox: Original vs. Matched Groups

Variable	Original	Group 1	Group 2	Group 3
Original loan amount (*1000)	237	243	245	240
FICO score	667	674	673	674
Original term	337	342	342	334
Issuance balance	235,195	241,585	242,440	237,518
Original LTV	71.36%	71.80%	71.55%	70.49%
Combined LTV	80.58%	80.28%	80.66%	80.09%
Original appraisal value (*1000)	347,190	349,681	349,977	353,142
Current interest rate	7.58	7.34	7.44	7.42
D_Second lien	27.83%	28.24%	27.98%	26.41%
D_subprime	26.57%	23.54%	24.15%	23.36%
D_Heloc	1.01%	0.22%	0.18%	1.10%
D_Interest only loan	19.57%	19.72%	19.6%	16.87%
D_FRM	41.33%	40.90%	41.55%	42.56%
D_Prepayment penalty	43.60%	44.33%	43.96%	41.01%
D_Condo	8.57%	9.97%	9.55%	8.20%
D_Single family	73.24%	75.40%	74.48%	74.37%
D_Multifamily	1.12%	0.86%	0.88%	1.06%
D_Option_ARM	5.51%	5.95%	6.12%	4.75%
D_Purchase loan	41.77%	47.14%	47.23%	41.01%
D_Refinance loan	47.83%	52.86%	50.79%	49.81%
D_Owner occupied loan	80.94%	83.56%	82.99%	85.88%
D_Investment loan	11.34%	15.53%	15.05%	10.39%
D_Full documentation	33.41%	34.72%	33.65%	32.95%
D_Low/No documentaion	36.57%	40.67%	41.99%	33.18%
Sample Size (*1000)	12,633	5,103	4,779	4,602

Note:

This table presents the summary statistics of BlackBox Analytics (BBX) dataset. Three groups of data are compared: the original BBX dataset (Original), the total matched sample drawn from pure statistical hard matching (Group1), statistical hard matching combined with machine learning techniques (Group2), and propensity score matching (Group 3). Those datasets include loans originated over the period 2003–2010; only loans with original loan amount less than $10 million are included in the sample. The variables with "D_" represent dummies. The details and explanations of the variables are presented in Table 5.1.

5.5.3 Bootstrapping analysis

Option based theoretical and empirical models for mortgage default analysis have been well developed during the past two decades (see, for example, Kau et al., 1992; Kau and Keenan 1999; Deng et al., 1996, 2000), and they have increased in realism and sophistication in the past decade (see Ambrose et al., 2001; Deng and Gabriel 2006 as two examples). In this analysis, we follow the literature as well as our first paper and estimate the accuracy of the matched sample using bootstrapping logistic analysis on the outcome of default.

In statistics, bootstrapping is a method for assigning measures of accuracy to sample estimates (Efron, 1993). This technique allows estimation of the sampling distribution of almost any statistic using only very simple methods (Varian, 2005). The basic idea of bootstrapping is that the sample we have collected is often the best guess we have as to the shape of the population from which the sample was taken.[62] Bootstrap offers to provide a way to simulate repeated observations from an unknown population using the obtained sample as a basis.

As an example, assume that we are interested in the average (or mean) height of people worldwide. We cannot measure all the people in the global population, so instead we sample only a part of it, and measure that. Assume the sample is of size N, and then we measure the heights of N individuals. From that

[62] For instance, a sample of observations with two peaks in its histogram would not be well approximated by a Gaussian or normal bell curve, which has only one peak. Therefore, instead of assuming a mathematical shape (like the normal curve or some other) for the population, we instead use the shape of the sample.

single sample, only one value of the mean can be obtained. In order to reason about the population, we need some sense of the variability of the mean that we have computed.

To use the simplest bootstrap technique, I take our original data set of 1/2 heights, and, using Stata, make a new sample (called a bootstrap sample) that is also half size of the original sample. The new sample is taken from the original one using sampling with replacement so it is not identical with the original "real" sample. I repeat this step 2000 times, and for each of these bootstrap samples the logistic test is conducted on the outcome of default on the same variables used in Table 3.2, Chapter 3 and the estimates of the coefficients (each estimate is called bootstrap estimate) are retained. A histogram of bootstrap estimates is now presented. This provides an estimate of the shape of the distribution of the mean from which questions about how much the mean varies can be answered. These estimates also show the amount of increase in the predicted log odds of default = 1 that would be predicted by a one unit increase in the predictor, holding all other predictors constant. With the estimated results, we calculate the predicted log odds of default = 1 with the actual ones, and obtain the probability of predicting correctly. Finally the results of accuracy predicted probability among the three matched groups are compared.

As a result, under the bootstrapping analysis, the probability for predicting the outcome of default in Group 2 (with machine learning) is around 94%, compared with 89% in Group 1 (pure statistical hard matching) and 75% in Group

3 (propensity score matching). Therefore, the bootstrapping analyses support that among the three approaches, statistical hard matching with machine learning performs the best, followed by pure statistical hard matching. The propensity score matching is considered to predict the outcomes least correctly.

In summary, the representativeness analyses above all confirm that the machine learning approach did a better job in dealing with selection bias and misclassification relative to the traditional approaches used in social science, such as pure statistical hard matching and propensity score matching. However, it is also shown that the performance is statistical hard matching, while not the best, is acceptable when there is no alternatives. Propensity score matching, although well packaged in various programs, should be used more carefully.

5.6 Summary

This study compares various data linkage approaches to deal with probabilistic data linkage in real estate studies. Previous analyses mainly focus on statistical hard matching, which identifies common covariates among different datasets and links these data using the common covariates, or propensity score matching (PSM), which refers to the pairing of treatment and control units with similar values on the propensity score, and possibly other covariates, and the discarding of all unmatched units (Rubin, 2001). Both statistical hard matching and PSM are commonly used in real estate studies, but both methods are criticized in dealing with selection bias and misclassification errors, as well as other limits.

As such, exploring other linking approaches and comparing with current linking methods in real estate studies is worthy, for the purpose of more reliable results with the matches sample. Machine learning, commonly used in the field of computer science, medicine, statistics, etc., is a key technique that exploits the nature of the dataset, e.g., the underlying patterns and relationship of variables. Classification algorithms such as Naive Bayes and Decision Tree work in the following way: the classifiers first learn the underlying pattern (term as model) from a set of labeled data (term as training set), and then apply the model to the unseen data and predict the label (matched or non-matched in this analysis) for this predicting set. Machine learning techniques can identify the potential patterns and especially some unobvious patterns that even human expert cannot easily figure out from data. Such knowledge makes the decision in data linkage wisely, instead of selecting randomly. In machine learning, the data linkage problem can be formulated as a binary classification problem, that is, judging whether two records from different datasets belong to the same entry or not. This approach may act as an alternative to link multiple datasets and deal with selection bias and misclassification errors.

Hence, the BBX into HMDA data are linked with the common covariates among these two datasets, using the three approaches: pure statistical hard matching as in the literature, statistical hard matching combined with machine learning techniques, and propensity score matching. Three groups of linked results from these approaches are generated and compared accordingly. As a result, under the pure statistical hard matching and the matching with machine learning, there are 2.5 million (20% of the original BBX) one-to-one exact matches. After eliminating observations with duplicate BBX id in multiple matches under pure statistical hard matching, there remain 2.6 million "actual" matches left, which consists of 21% of the original BBX data. With machine learning techniques, it is shown that the model from Decision Tree Classifier better fits the data situation, with higher estimation score than Naïve Bayes Classifier. The high probability of correct linking (above 80%) suggests that the one-to-one matched sample can be considered as true match exempted from misclassification issue. The trained model further identifies around 18% matches with unique BBX ids from the original multiple matches. Under propensity score matching, it obtains only 26 thousand matches (less than 1% of the original BBX data) with exactly same propensity scores, but generates another 4.6 million matches (round 36% of the original BBX data) when the match is based on approximate propensity scores. In summary, by comparing the number of linkages under the three approaches, it is observed that using statistical hard matching (Group 1) obtains slightly more linking records than using statistical hard matching with machine learning (Group 2), followed by propensity score matching (Group 3).

To confirm the absence of sample selection problem, several checks are conducted to compare the three matched groups with the original BBX and HMDA data, including the approaches frequently used by other economic studies such as distributions of the key variables, and summary statistics comparison. Firstly, the distributions of the key variables are examined; for those continuous attributes, the kernel density distributions are analyzed, while and frequency plots for categorical attributes are shown. Secondly, the summary statistics of crucial variables from the matched and the original samples are conducted and compared. The bootstrapping approach is also used to estimate the outcome of default based on the key variables from the linked groups. Overall, the findings imply that statistical hard matching with machine learning approach did a better job in dealing with selection bias and misclassification relative to the traditional approaches used in social science, such as pure statistical hard matching and propensity score matching. What's more, by repeating 2000-time logistic regression analyses in bootstrapping approach, we find that the probability for predicting the outcome of default in the matched sample drawn from machine learning is higher than that from other approaches, which further supports the advantage of machine learning approach towards others.

The total matched dataset under machine learning approach provides us good opportunities to conduct innovative analysis, by examining racial, ethnic, gender, and income differences in mortgage lending, controlling for both the risk profile of the mortgage and the characteristics of the neighborhood where the property is located. The machine learning approach used in data linkage procedure is beneficial to not only real estate studies, but also any data matching issues trying to deal with sample selection problems and misclassification issues.

However, it is also shown that due to the complexity of the machine learning techniques, it is sometimes hard to apply and explain, especially to readers from fields other than computer science. In comparison, the performance of statistical hard matching, while not the best, is generally acceptable when there are no alternatives. Propensity score matching, although commonly accepted in statistics and social science and well packaged in various programs, should be used more carefully.

In summary, shown in Section 2.5, Chapter 2, as with any linkage, the quality of the match is limited to the quality of the original data se well as the ability of the vector covariates to distinguish uniqueness among the two populations. An error-free "match" is not guaranteed. Potential for mismatch because of recording errors, different recording conventions, or changes to information over time may be occurred. Although the match-merge process is designed to control matching error, the conclusions to be drawn from any match should ultimately rely on the quality of each data repository.

Chapter 6 Conclusions

This research firstly aims to investigate the unique risk patterns of borrowers, especially investors and their behaviors in the U.S. condominium (condo) loan market in the early 2000s, which have been overlooked in understanding the financial crisis. Secondly, it examines the impact of neighborhood foreclosure concentration on individual borrower's delinquency decision, which is crucial to mortgage default risk management, pricing and underwriting. Finally, it compares and discusses the advantages and disadvantages of various approaches in dealing with data linkage issues in real estate studies, aiming to help provide certain implications for future real estate research. Three studies were conducted accordingly and results found that, using U.S. condominium market as a natural experiment, mortgage borrowers, especially investors play a significant role in understanding the current financial crisis. In addition, findings showed that neighborhood foreclosure concentration increases borrowers' default option exercise during the study period, but the impacts differ in different regimes and across different borrower groups. The comparison of various data linkage approaches and results reveal that statistical hard matching with machine learning approach did a better job in dealing with selection bias and misclassification relative to the traditional approaches used in social science, such as pure statistical hard matching and propensity score matching. In this chapter, I first briefly review the research and then highlight the contributions of the research. Finally, I summarize the limitations and future research.

Chapter 6

6.1 Review of the Research

The U.S. market have experienced the longest periods of housing market booms in history. The surging house prices have resulted in lower mortgage interest rates, lower down payment criteria, more financing alternatives and more relaxed lending standards, compared with previous stage. However, the great expansion of mortgage lending has led to great credit risks, which has resulted in the substantial surge in the residential mortgage delinquencies, followed by the collapse of the house price boom in the U.S. housing market and the recent financial crisis. Such collapse in the values of mortgages further brings a substantial increase in foreclosures and large decline in house prices. As a result, the fast increasing foreclosures and house price drops recursively lead to the increasingly worse housing market.

The current mortgage crisis and the following disasters on the financial market induce various discussions on the possible triggers of this crisis, among academia and practitioners. Literature on the possible triggers of the current financial crisis, including innovation in mortgage products, fast growth of securitization, the market players' wisdom and borrower behavior helps to get a better understanding on the crisis. However, borrower behavior especially investor behavior in mortgage choices is crucial in triggering the current financial crisis but largely overlooked in previous literature, due to the data limits and identification issues of investors from consumers in the housing market.

This research documents the unique risk patterns of borrowers, especially investors and their behaviors in the U.S. condominium (condo) loan market in the early 2000s, which have been overlooked in understanding the financial crisis. This analysis addresses question regarding the default probability of the condo loans relative to commonly discussed single-family mortgages conditioning on various loan and borrower characteristics, macroeconomic conditions and so on. In addition, several competing explanations for the observed evidence on the faster default growth in the condo loan market are examined, such as the unique characteristics of the condo home markets which is the unobserved heterogeneity issue, the lender (supply side) effect, and the borrower (demand side) effect. The following questions are investigated:

1) Do condo loans differ from single-family loans with respect to default patterns?
2) If yes, what is the driving factor of the unique default pattern in the condo loan market?
3) In a neighborhood with condo loan defaults, does impact of condo loan defaults resulting from risky borrowers have negative spillover effects on the neighboring single-family loans? Or do early condo defaults predict the neighboring single family subprime market's subsequent default rate?

If the third assumption that defaults and thus foreclosures have spillover effects on nearby borrowers' default probabilities holds, the great influences of neighborhood defaults and foreclosures are worth deeply studied in better understanding the intrinsic mechanism of the crisis and finding the solutions. Since recently, the impact of neighborhood foreclosure concentration on individual borrower's delinquency probability has been increasingly emphasized and studied. There is a great amount of evidence that foreclosures can have great influences on neighborhoods, from the view of house price decline in neighborhoods, rise in violent crime and thefts and thus the instability of the community, acceleration of racial transition, children performance, and emotional and physical impact on people. However, the impact of foreclosure concentration on the borrower's sensitivity to negative equity, i.e., to a certain extent the changing attitude of borrowers towards default option exercise, has not been fully discussed. Thus whether the information effect or foreclosure contagion effect dominates neighborhood foreclosure concentration impact on nearby borrowers' delinquency decision is an open question.

As such, I examine the impact of neighborhood foreclosure concentration on individual borrower's delinquency probability. I also estimate foreclosure concentration on the borrower's sensitivity to negative equity, which is to a certain extent the changing attitude of borrowers towards default option exercise. Accordingly, the following questions are examined:

1) Does concentrated foreclosure increase or decrease the probability or the attitude of the geographically neighboring borrowers to make their default decision?
2) Will these increases or decreases differ in different regimes and across different borrower groups?

When digging into the research about borrower and investor behavior and foreclosure concentration, it is shown that in most cases, while each of these datasets provides certain information, these datasets lack a significant amount of information due to the constraints of data sources, thus no single source of data has all of the information required for certain undertaking. Given these challenges and data limitations, in the absence of the ideal data source with complete and necessary information, there is a necessity to link records in two or more separate but intrinsically correlated data sets in case when exact matching of individual records is not possible due to confidentiality restrictions on the data available, to overcome the limitations of existing data sources, thereby enhancing the application of datasets. However, current linking approaches such as statistical hard matching and propensity score matching are observed to have potential shortages in linking multiple datasets.

Therefore, more advanced and well-developed technique in the field of computer science, machine learning, is applied to deal with multiple matches and compared with other approaches such as pure statistical hard matching and propensity score matching. Research questions are shown accordingly:

1) By comparing various linking approaches, what is the appropriate one to link multiple datasets in real estate studies, especially mortgage studies, when there are no unique identifiers?
2) Among the linkage approaches, how can we minimize the selection bias and identification errors?

The findings in the first research confirm the notion that borrower behavior, especially investor behavior, plays a significant role in understanding the current financial crisis. First of all, the results document that there is a sharp increase in condo loan defaults relative to single-family loan defaults over the years. Condo loan default rate also grows at a faster rate, even compared with subprime loans. What's more, it is shown that the leading factor of the unique default pattern in the condo loan market is due to inherently riskier loan borrowers in the condo loan market, compared with single family loan market: investment-purchase condo loans are much more likely to default compared to other condo loans, and the effect is strengthened when the option to default is more in the money. Last but not least, not only should condo loans default earlier compared with single-family loans originated in the same cohort, but also the earlier condo loan defaults prompt more defaults in the single-family sector in the same area afterwards.

The results on foreclosure concentration impacts reveal that on average neighborhood foreclosure concentration enhances borrowers' default option exercise during the study period – borrowers are more willing to enter into default when there are intense foreclosures in the neighborhood. However, interestingly, the impact of foreclosure concentration varies in different regimes: before 2007, higher neighborhood foreclosure intensity is associated with reduced borrower sensitivity; entering into the crisis period (2007-2011), the impact turns from negative to positive; and post 2012, the impact becomes insignificant. The net impact of neighborhood foreclosure concentration on borrowers' sensitivity to negative equity also varies across different borrower groups.

The outcomes of applying and comparing various data linkage approaches show that, in general, statistical hard matching obtains slightly more linking records than using statistical hard matching with machine learning, and much more than using propensity score matching. After that, several representativeness analysis results on the linked groups are presented, including examining the distributions of the key variables (by looking at kernel density distributions for continuous attributes and frequency plots for categorical attributes), comparing the summary statistics of the matched and original samples, and conducting the bootstrapping analysis on the outcome of default based on the key variables from both datasets. Findings support that statistical hard matching with machine learning approach is a comparatively better approach in dealing with selection bias and misclassification, relative to the traditional approaches such as pure statistical hard matching and propensity score matching. Propensity score matching, although well packaged in various programs, should be used more carefully. The performance of statistical hard matching, while not the best, is generally acceptable when there are no alternatives.

These findings in the three studies have several implications. First, it implies that condo loans are inherently riskier than single-family loans. The evidence also suggests that investment-driven, riskier borrowers in the condo market are the most plausible driver for the observed default patterns in this market. What's more, investors are more responsive to market conditions in their default behavior. Given that real estate investors are more present in the condominium market, the observed default pattern in the condo loan market thus may be associated with the investor behavior. Second, the impact of neighborhood foreclosure concentration on borrower default behavior is not limited to the contagion effect. It is actually shown that sometimes the impact can be on the opposite direction – foreclosures can discourage borrower's delinquency if borrowers take foreclosures as a signal of how lenders will deal with delinquencies. This information effect can dominate the contagion effect during the market boom. From this perspective, borrowers are strategic in their default decisions. Credit risk modelers thus should take this game feature of mortgage default into consideration to achieve better understanding and estimation of mortgage default risk.

Overall, the results of this research answer the questions that I attempted to investigate. Although it is impossible to stop the influences of the financial crisis, prevent investors from leading to worse market, or solve the problems of data linkage through this research, it does provide a better understanding of those problems. The contributions of the current research are summarized in the next section.

6.2 Potential Contributions

This research enriches the literature, provides alternative explanations for real-life problems and sheds lights on policies that are helpful to address these problems.

First, with respect to the literature contributions, the study in Chapter 3 is of great significance. This study is the first to document a strong, robust and economically important default pattern in the much ignored condominium loan market. Previously almost all mortgage studies focus on single-family loan market. Specifically, the loan origination growth rate and default pattern in the condo market are comparable to the subprime mortgage market (Demyanyk and Van Hemert, 2011). Condominium borrowers, the group of which is less studied, are unlikely the low credit quality borrowers who default because they cannot afford to pay or refinance their mortgages as house prices start to decline, tend to have higher FICO scores, use subprime mortgages less frequently, and on average are charged a lower interest rate. These unique characters of condo borrowers may reveal distinct default behavior compared with single-family borrowers. Therefore, the condominium loan market, given the characteristics, provides a unique opportunity to identify and analyze the investor behavior.

Second, the new empirical evidence from the influences of borrower behaviors, as revealed in this research, adds to the understanding of the economic channels that explain the financial crisis. The findings in Chapter 3 complement the demand-side view by providing evidence that investor behavior, as manifested in the condo market's loan default pattern in our context, play an important role in explaining mortgage defaults in the crisis. It is shown that investment-purchase condo loans not only drive the observed condo loan market default pattern in triggering more defaults, but condo defaults also prompt more defaults of single family subprime mortgages at the same location, through the channel that foreclosures on the defaulted properties depress neighboring house prices. The findings that condo borrowers, especially investors, are riskier also suggest that lenders need to exercise more scrutiny in their lending practice in the condominium mortgage market. From a public policy point of view, it is found that simply requiring more skin-in-the-game regulations for lenders and lower LTV for the borrowers under the Dodd-Frank law is only a partial solution from avoiding a similar crisis in the future. Therefore, this thesis's attempt to study the characteristics and delinquency probabilities of loans from borrowers' perspective is not only academically meaningful, but also is important in explaining what we have experienced in the recent crisis.

Third, the foreclosure concentration study in this research is among the few direct analyses of the neighborhood foreclosure concentration. Identifying and understanding concentration effects in foreclosures helps provide a better understanding of how and why such crises spread. Traditional studies of borrower decision mainly focus on mortgage borrowers' own socio-economic status such as the borrower's FICO score, income constraint, and equity position. However, to place borrowers into social networks to understand their default decisions are crucial to mortgage default risk management, pricing and underwriting. Comparing to existing studies, this work takes a novel approach to not only assess the impact of neighborhood foreclosure concentration on individual borrower's delinquency probability but also estimate the impact of foreclosure concentration on the borrower's sensitivity to negative equity, which is to a certain extent the changing attitude of borrowers towards default option exercise. The presence of foreclosure concentration is relevant to policy makers concerned with mitigating the spread of home foreclosures. The finding that peer behavior indeed has great influence on borrower's actual default choice indicate that those default models to predict the borrower's default risks should incorporate such network effects.

Fourth, from a policy perspective, the findings about the impact of foreclosure concentration on borrower's delinquency decision have great policy implications. The results show that increased delinquencies result in more foreclosures, and concentrated foreclosure further result in even more delinquencies. Thus, mortgage default, especially during the crisis, can be self-enforcing in certain neighborhoods. Considering the great financial and social impacts of mortgage defaults, and the potential recursive enforcing of foreclosures in the same neighborhoods, these findings call for the government's timely intervention to reduce foreclosure, not only for current situation, but also to break the loop and stop the foreclosure cascade in the future.

Fifth, the application of data linkage in the fifth chapter is potentially useful in filling in additional or missing information, by adding in extra attributes. With more complete information on population units more complex research questions can be further addressed. Linking multiple datasets might be a way of checking accuracy and reliability of survey or administrative data or vice versa; one can assess whether the sample survey data are producing reliable inferences using some population administrative datasets to assess the representativeness of the sample data. Last, linking records helps enhance data quality, by providing more information for people to understand the non-response or non-report side of the current data.

The data linkage study also extends the literature by comparatively analyzing different linking methods on multiple mortgage datasets to help better understand the advantages and potential limits of each method, as well as trying to overcome the selection bias and misclassification issues. In particular, this study systematically compares the commonly used approaches for probabilistic linkage and applies advanced techniques from computer science field to try to solve the selection bias problems in linking process, by letting the computer to exploit the nature of the dataset, e.g., the underlying patterns and relationship of variables based on both common covariates and the rest covariates. These attempts help provide a creative way to other authors who also need to link multiple data sources with no unique identifiers and conduct deep analysis based on a representative matching dataset.

In summary, this research achieves its objectives. The findings answer my research questions and are meaningful to address the targeted research problems. Therefore, the significance of this research, as mentioned in the introduction chapter, has been realized.

6.3 Limitations and Future Research

In this section, all of the limitations of this research and their reasons are listed, followed by a description of the future research that I intend to conduct.

Chapter 6

The first essay is one of the first studies in U.S. to document the unique risk patterns of borrowers, especially investors and their behaviors in the U.S. condominium (condo) loan market in the early 2000s. Results in this essay imply that condo loan market is an important channel to understand the cause and transmission mechanism of the recent financial crisis especially from the perspective of borrowers and investors' behavior. However, the evidence in this study only provides the first step in studying the cause and aggregate implications of the condo loan defaults from borrowers' perspective. Future research can be extended to better understand the role of borrowers, especially investors in that market in fueling and potentially exacerbating the crisis. I will attempt to study the changing behaviors of investors through the whole default process, across the financial crisis and in different regions, rather than focusing on the final delinquency outcome.

Throughout the first study, the most important role is the investor channel: the price run-up in the earlier years of the decade attracts more investors who are also more likely to make a pure economic decision of deciding to default soon after a significant price drop. In order to examine the investor behaviors, this study uses the U.S. condominium market as the natural experiment, since results show that real estate investors are more present in the condominium market. Therefore, the observed default pattern in the condo loan market thus may be associated with the investor behavior. However, the identification and analyses of investors are constrained by the data: whether the borrower's purpose of buying the property as a homeowner or investor is self-reported by the borrower and recorded by the mortgage lender. Thus in the attempt to obtain a better mortgage, those borrowers may not report their actual purpose. Although some attributes in the data can be regarded as the criteria for judging the investor or homeowner, such as the FICO score, this is not the direct way to identify the "actual" investors. In the future the study will be better supported if a comparably trustable reported or discovered investor data is applied. Nevertheless, note that the current study does contain a sample that represents the relative best information about investor, and it is the best data that can be found for studying the investor behavior till now. In addition, although some borrower characteristics are obtained, the first essay still lacks of other demographic characteristics of condo borrowers, such as education, their preferences in living areas, their revenues and expenditures, etc. For example, Chinco and Mayer (2014) study the local and out-of-town long distance investors, which is quite interesting to me. Unfortunately, the current information is the best that I can obtain. In the future research I would try to find this kind of information, and further study the condo investor behaviors.

The research in Chapter 4 presents rich findings about foreclosure concentration effects on neighborhood borrowers' default option exercise, through different regimes, and across different group of people. The evidence shows that foreclosures can induce nearby mortgage borrowers to exercise their default option more ruthlessly, which is especially prominent during a downturn of the housing market. However, there are also some other ways that can generate foreclosure contagion, such as observational learning, herding and so on. More specifically, observational learning suggests that homeowners update their beliefs about the value of their homes when they receive signals about house price trend (Agarwal et al., 2012). Foreclosures in one's neighborhood send out a public signal of a declining property market. Based on such a signal, nearby homeowners will adjust their valuation downward, causing an observed negative impact of nearby foreclosure on property values. Such downward adjustment in valuation apparently increase the probability of default as borrowers default their mortgage loans mainly because the value of the property is lower than the mortgage loan balance. Another channel of foreclosure contagion is through herding. Homeowners are easily persuaded to follow the herd to strategically default their mortgage loan (Seiler et al., 2014). Extending this herding rational to mortgage borrower's delinquency decision, someone who resides in a neighborhood with concentrated foreclosures is exposed to the influence of her neighbors and thus is more likely to exercise her default option when she sees many foreclosure signs in her neighborhood. In addition, there might be moral issue that seeing many neighbors have done so might have changed some borrowers' view. However, among these mechanisms through which the foreclosure contagion affects the borrower's default option exercise, it is not clear what the direct mechanism is in this study. In order to better understand the foreclosure contagion effect, future research should try to establish the exact mechanism of the foreclosure contagion discovered in this paper, and assess the relative roles of observational learning, herding and other channels in generating such foreclosure contagion.

Furthermore, the fourth essay only focuses on LA MSAs, which is relatively a restricted sample. The reason is that I want to have a more homogeneous sample without the trouble of location differences. But it is also interesting to look at those areas with distinct foreclosure phenomenon. For example, Arizona is among the leading 6 states that have the greatest foreclosure concentration; however, its foreclosure concentration drops quickly afterwards. In the meantime, Illinois and Ohio did not appear in the leading states in 2011, but these two states become the top states that lead the nation's foreclosure activity in 2012. Seeing how the changing foreclosure concentration in these states will influence the attitudes would be interesting and it might help understand the foreclosure concentration effects from a different angle. This direction might be my next step in trying to better understand the foreclosure concentration impacts in U.S.

Finally, the results in Chapter 5 suggest that statistical hard matching with machine learning approach performs better in dealing with selection bias and misclassification, relative to the traditional approaches such as pure statistical hard matching and propensity score matching. However, the results from pure statistical hard matching and the one with machine learning do not differ much in general; the performance of statistical hard matching, while not the best, is generally acceptable when there are no alternatives. As we know that as with any linkage, the quality of the match is limited to the quality of the original data se well as the ability of the vector covariates to distinguish uniqueness among the two populations. Therefore, it might be possible that the similarity of the performances between these two groups might be partially due to the specific characteristics of the data. Therefore, in future research it will be good to apply and compare these approaches to more general data in real estate studies, to check the performances of the resulted groups. In addition, it should be noted that due to the complexity of the machine learning techniques, it is difficult to apply and explain these techniques, especially to the readers from fields other than computer science. In order to make this approach better understood and applied, future research might focus on the simplified machine learning which is more suitable to real estate studies.

Bibliography

Agarwal, Sumit. 2007. "The Impact of Homeowners' Housing Wealth Misestimation on Consumption and Saving Decisions," Real Estate Economics 35 (2): 135–154.

Agarwal, S., Ambrose, B., Chomsisengphet, S., Sanders, A. 2012. The neighbor's mortgage: does living in a subprime neighborhood impact your probability of default? Real Estate Economics 40(1), 1-22.

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., Evanoff, D. 2013. Predatory Lending and the Subprime Crisis. Forthcoming, Journal of Financial Economics.

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., Evanoff, D. 2011. The role of securitization in mortgage renegotiation. Journal of Financial Economics 102(3), 559-578.

Agarwal, S., Ben-David, I., Yao, V. 2012. Appraisal bias: evidence from the residential real-estate market. Working paper. The Ohio State University.

Agarwal, S., Chang, Y., Yavas, A. 2012. Adverse selection in mortgage securitization. Journal of Financial Economics 105(3), 640-660.

Alexander, W. P., Grimshaw, S. D., McQueen, G. R., & Slade, B. A. 2002. Some Loans Are More Equal than Others: Third–Party Originations and Defaults in the Subprime Mortgage Industry. Real Estate Economics, 30(4), 667-697.

Bibliography

Alpaydin, Ethem. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning), MIT Press.

Ambrose, B.W., Capone, C.A., Deng, Y. 2001. Optimal put exercise: an empirical examination of conditions for mortgage foreclosure. Journal of Real Estate Finance and Economics 23, 213-234.

Amromin, G., Huang, J., Sialm, C., Zhong, E. 2013. Complex mortgages, unpublished working paper, Federal Reserve Bank of Chicago, University of Texas at Austin, University of Texas at Austin, National Bureau of Economic Research (NBER), University of Wisconsin-Madison.

An, X., Deng, Y., Gabriel, S. 2011. Asymmetric information, adverse selection, and the pricing of CMBS. Journal of Financial Economics 100 (2), 304-325.

An, X., Deng, Y., Rosenblatt, E., Yao, V.W. 2012. Model stability and the subprime mortgage crisis. Journal of Real Estate Finance and Economics 45 (3), 545-568.

An, M. Y., & Qi, Z. 2012. Competing Risks Models using Mortgage Duration Data under the ProportionalHazards Assumption. Journal of Real Estate Research, 34(1), 1-26.

Avery, Robert B., Kenneth P. Brevoort, and Glenn B. Canner. 2007. The 2006 HMDA Data, Federal Reserve Bulletin, Vol. 93. A73-A109.

Barlevy, G., Fisher, J. 2011. Mortgage choice and housing speculation. Chicago Fed working paper.

Bibliography

Been, V., Ellen, I. G., Schwartz, A. E., Stiefel, L., & Weinstein, M. 2011. Does losing your home mean losing your school?: Effects of foreclosures on the school mobility of children. Regional Science and Urban Economics, 41(4), 407-414.

Ben-David, I. 2011. Financial constraints and inflated home prices during the real-estate Boom. American Economic Journal: Applied Economics 102(3), 559-578.

Ben-David, I. 2012. High leverage and willingness-to-pay: evidence from the residential housing market. Working paper. The Ohio State University.

Baxter, V., & Lauria, M. 2000. Residential mortgage foreclosure and neighborhood change. Housing Policy Debate, 11(3), 675-699.

Bikhchandani, S., Hirshleifer, D., & Welch, I. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. The Journal of Economic Perspectives, 151-170.

Blanchette, C. M., DeKoven, M., De, A. P., & Roberts, M. 2013. Probabilistic data linkage: a case study of comparative effectiveness in COPD. Drugs in context, 2013.

Campbell, T. S., Dietrich, J. K. 1983. The determinants of default on insured conventional residential mortgage loans. The Journal of Finance, 38(5), 1569-1581.

Campbell, J.Y., Gilio, S., Pathak, P. 2011, Forced Sales and House Prices, American Economic Review, 101 (5): 2108-2131.

Bibliography

Campbell, J. Y., Ramadorai, T., & Ranish, B. 2014. The Impact of Regulation on Mortgage Risk: Evidence from India, American Economic Journal: Economic Policy, forthcoming.

Capone, C. A. 2001. Introduction to the special issue on mortgage modeling. The Journal of Real Estate Finance and Economics, 23(2), 131-137.

Case, K. E., & Shiller, R. J. 1989. The behavior of home buyers in boom and post-boom markets.

Case, K. E., Shiller, R. J., & Thompson, A. 2014. What Have They Been Thinking? Homebuyer Behavior in Hot and Cold Markets,, unpublished working paper, Wellesley College, Yale University, McGraw-Hill Construction, Social Science Research Network.

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. 2003. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. American journal of epidemiology, 158(3), 280-287.

Cheng, I., S. Raina, and W. Xiong. 2013. Wall Street and the housing bubble. Working paper.

Chiang, T. C., & Zheng, D. 2010. An empirical analysis of herd behavior in global stock markets. Journal of Banking & Finance, 34(8), 1911-1921.

Chinco, A., & Mayer, C. 2014. Misinformed speculators and mispricing in the housing market, unpublished working paper, University of Illinois Urbana-Champaign College of

Business, Columbia Business School and NBER, National Bureau of Economic Research (No. w19817).

C. Clapp, J., Deng, Y., An, X. 2006. Unobserved heterogeneity in models of competing mortgage termination risks. Real Estate Economics, 34 (2), 243-273.

Clauretie, T.M., Jameson, M. 1995. Residential loan renegotiation: theory and evidence. Journal of Real Estate Research 10(2), 153-162.

Clauretie, T. M., & Sirmans, G. S. 2003. Real Estate Finance: Theory and Practice (4th ed.) Mason, OH: Thomas Learning.

Cordell, L., Dynan, K., Lehnert, A., Liang, N., Mauskopf, E. 2009. Designing loan modifications to address the mortgage crisis and the making home affordable program. Federal Reserve Board working paper.

Cooper, W. S., and M. E. Maron. 1978. "Foundations of Probabilistic and Utility-Theoretic Indexing", Journal of the Association for Computing Machinery, 25, pp. 67-80.

Crews Cutts, A., Merrill, W.A. 2008. Interventions in mortgage default: policies and practices to prevent home loss and lower costs. Freddie Mac working paper.

Daneshvary, N., Clauretie, T. M., & Kader, A. 2011. Short-term own-price and spillover effects of distressed residential properties: The case of a housing crash. Journal of Real Estate Research, 33(2), 179-207.

Danilo, Croce, Basili Roberto. 2010. Decision tree algorithm short Weka tutorial.

Dehejia, Rajeev H, Wahba, Sadek. 2002. Propensity score-matching methods for nonexperimental causal studies, The review of Economics and Statistics, 84 (1), 151-161.

Demyanyk, Y., Van Hemert, O.. 2011. Understanding the subprime mortgage crisis. Review of Financial Studies 24, 1848-1880.

Deng, Y., Quigley, J.M., Van Order, R.. 1996. Mortgage default and low downpayment loans: the costs of public subsidy. Regional Science and Urban Economics 26(3-4), 263–287.

Deng, Y. 1997. Mortgage termination: An empirical hazard model with a stochastic term structure. The Journal of Real Estate Finance and Economics, 14(3), 309-331.

Deng, Y., Quigley, J.M., Van Order, R. 2000. Mortgage terminations, heterogeneity and the exercise of mortgage options. Econometrica 68(2), 275–307.

Deng, Y., Quigley, J.M. 2002. Woodhead Behavior and the Pricing of Residential Mortgages. SSRN working paper.

Deng, Y., Pavlov, A. D., & Yang, L. 2005. Spatial heterogeneity in mortgage terminations by refinance, sale and default. Real Estate Economics, 33(4), 739-764.

Deng, Y., Gabriel, S.A. 2006. Risk-based pricing and the enhancement of mortgage credit availability among underserved and higher credit-risk populations. Journal of Money, Credit and Banking 38(6), 1431-1460.

Ding, L., & Quercia, R. G. 2010. Neighborhood Subprime Lending and the Performance of Community Reinvestment Mortgages. Journal of Real Estate Research, 32(3), 341-376.

Ding, L., Quercia, R. G., Li, W., & Ratcliffe, J. 2011. Risky Borrowers or Risky Mortgages Disaggregating Effects Using Propensity Score Models. Journal of Real Estate Research, 33(2), 245-277.

D’Orazio, M., Di Zio, M., & Scanu, M. 2006. Statistical Matching: Theory and Practice. Chichester: Wiley.

Duchin, R., Sosyura, D. 2010. TARP consequences: lending and risk taking, unpublished working paper, University of Washington, University of Michigan.

Dunn, H L. 1946. Record Linkage, American journal of public health and the nation's health, 36(12): 1412-1416.

Efron, B.; Tibshirani, R. 1993. An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC.

Elul, R. 2011. Securitization and mortgage default. Available at SSRN 1786317.

Fan, Wei, Davidson, Ian, Zadrozny, Bianca, and Yu, Philip S. 2005. An Improved Categorization of Classifier’s Sensitivity on Sample Selection Bias, Data Mining, Fifth IEEE International Conference.

Fellegi and Sunter. 1969. A Theory for Record Linkage, Journal of the American Statistical Association, Vol. 64, No. 328.

Ferreira, Fernando, Gyourko, Joseph. 2011. Anatomy of the Beginning of the Housing Boom : U.S. Neighborhoods and Metropolitan Areas, 1993-2009, Working Paper, The Wharton School, University of Pennsylvania & NBER.

Foote, Chris, Kristopher S. Gerardi and Paul S. Willen. 2008. Negative Equity and Foreclosure: Theory and Evidence. Journal of Urban Economics 64(2): 234–245.

Foster, C., & VANORDER, R. 1984. An option-based model of mortgage default. Housing Finance Review, 3(4), 351-372.

Foster, C., & Order, R. 1985. FHA terminations: A prelude to rational mortgage pricing. Real Estate Economics, 13(3), 273-291.

Fraeman, K. 2010. An introduction to implementing propensity score matching with SAS®. Bethesda, MD: United BioSource Corporation.

Frame, W.S. 2010. Estimating the Effect of Mortgage Foreclosures on Nearby Property Values: A Critical Review of the Literature. Working paper. Federal Reserve Bank of Atlanta.

Fu, Y., Qian, W. 2012. Speculators and price overreaction in the housing market. Working paper. National University of Singapore.

Fu, Y., Qian, W., Yeung, B. 2012. Housing Market Speculators and Transaction tax. NBER working paper w19400.

Galindo, J., and P. Tamayo. 2000. "Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications." Computational Economics 15.1-2: 107-143.

Gangel, M., Seiler, M. J., & Collins, A. 2013. Exploring the foreclosure contagion effect using agent-based modeling. The Journal of Real Estate Finance and Economics, 46(2), 339-354.

Garmaise, Mark. 2013a. The Attractions and Perils of Flexible Mortgage Lending, forthcoming, Review of Financial Studies.

Garmaise, M. J. 2015. Borrower misreporting and loan performance, The Journal of Finance 70(1), 449-484.

Garmaise, Mark J. 2013b. Borrower Misreporting and Loan Performance, Journal of Finance, forthcoming.

Gerardi, K., Shapiro, A. H., & Willen, P. S. 2008. Subprime outcomes: Risky mortgages, homeownership experiences, and foreclosures (No. 07-15). Working paper series//Federal Reserve Bank of Boston.

Gerardi, K., & Willen, P. 2009. Subprime mortgages, foreclosures, and urban neighborhoods. The BE Journal of Economic Analysis & Policy, 9(3).

Getoor, L., Friedman, N., Koller, D., and Taskar, B. 2003. Learning Probabilistic Models for Link Structure, Journal Machine Learning Research, 3, 679-707.

Ghent, Andra C., Hernandez-Murillo, Ruben, and Owyang, Michael. 2011. Race, Redlining, and Subprime Loan Pricing, Working Paper, Federak Reserve Bank of St. Louis.

Ghent, Andra C. and Kudlyak, Marianna. 2011. Recourse and Residential Mortgage Default: Evidence from U.S. States. Review of Financial Studies 24(9): 3139-3186.

Good, I.J. 1950. Probability and the Weighing of Evidence, London, Charles Grin.

Goodstein, R., Hanouna, P., Ramirez, C. D., & Stahel, C. W. 2011. Are Foreclosures Contagious?. Available at SSRN 2024794.

Gorton, G.B., Pennacchi, G.G. 1995. Banks and loan sales: marketing nonmarketable assets. Journal of Monetary Economics 35, 389-411.

Gramlich, E. M. 2007. Subprime Mortgages: American's Latest Boom and Bust. Washington, DC: The Urban Institute Press.

Guiso, L., Sapienza, P., & Zingales, L. 2009. Moral and social constraints to strategic default on mortgages (No. w15145). National Bureau of Economic Research.

Guiso, L., Sapienza, P., & Zingales, L. 2013. The determinants of attitudes toward strategic default on mortgages. The Journal of Finance, 68(4), 1473-1515.

Gu, Lifang, Baxter, Rohan, Vickers, Deanne, and Rainsford, Chris. 2003. Record Linkage: Current Practice and Future Directions, CSIRO Mathematical and Information Sciences.

Hammill, B. G., Hernandez, A. F., Peterson, E. D., Fonarow, G. C., Schulman, K. A., & Curtis, L. H. 2009. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. American heart journal, 157(6), 995-1000.

Harding, John P., Rosenblatt, Eric and Yao, Vincent W. 2008. The Contagion Effect of Foreclosed Properties. Journal of Urban Economics 66(3): 164-178.

Haughwout, A., Lee, D., Tracy, J. S., & Van der Klaauw, W. 2011. Real Estate Investors, the Leverage Cycle, and the Housing Market Crisis, , Federal Reserve Bank of New York Staff Reports 514.

Haughwout, Andrew, Mayer, Christopher, and Tracy, Joseph. 2009. Subprime Mortgage Pricing: The Impact of Race, Ethnicity, and Gender on the Cost of Borrowing, Federal Reserve Bank of New York Staff Reports.

Haughwout, A.F., Peach, R.W., Tracy, J. 2008. Juvenile delinquent mortgages: bad credit or bad economy? Journal of Urban Economics 64(2), 246-257.

Heckman, J. J. 2005. The scientific model of causality, Sociological Methodology, 35(1), 1-98.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. Characterizing Selection Bias Using Experimental Data. Econometrica 66 (5): 1017–1098.

Hernandez-Murillo, Ruben, Sengupta, Rajdeep. 2012. The Effect of Neighborhood Spillovers on Mortgage Selection, Working Paper, Federak Reserve Bank of St. Louis.

Hosmer, David W.; Lemeshow, Stanley. 2000. Applied Logistic Regression (2nd ed.). Wiley.

Igan, D., Mishra, P., Tressel, T. 2011. A Fistful of Dollars: Lobbying and the Financial Crisis, , NBER Macroeconomics Annual, 26, w17076.

Immergluck, D., Smith, G. 2006a. The impact of single-family mortgage foreclosures on neighborhood crime. Housing Studies 21(6): 851-866.

Immergluck, D., and G. Smith. 2006. The External Costs of Foreclosure: The Impact of Single-Family Mortgage Foreclosures on Property Values. Housing Policy Debate, 17(1): 57-79.

Immergluck, D. 2008a. Community Response to the Foreclosure Crisis: Thoughts on Local Interventions (Community Affairs Discussion Paper No. 01-08).

Ioannides, Y. M. 2003. Interactive property valuations. Journal of Urban Economics, 53(1), 145-170.

Jaro, M. A. 1995. Probabilistic linkage of large public health data files. Statistics in medicine, 14(5-7), 491-498.

Jiang, W., Nelson, A. A., & Vytlacil, E. 2014 Liar's loan? Effects of origination channel and information falsification on mortgage delinquency,.Review of Economics and Statistics 96(1), 1-18.

Jiang, W., Nelson, A., Vytlacil, E. 2011b. Securitization and loan performance: a contrast of ex ante and ex post relations in the mortgage market. Working Paper. Columbia University.

Joffe, Marshall M., and Paul R. Rosenbaum. 1999. Invited commentary: propensity scores."American journal of epidemiology 150.4: 327-333.

Kain, John F., and John M. Quigley. 1972. Housing market discrimination, home-ownership, and savings behavior. The American Economic Review 62.3: 263-277.

Kau, J. B., & Keenan, D. C. 1995. An overview of the option-theoretic pricing of mortgages. Journal of Housing Research, 6(2), 217-244.

Kau, J.B., Keenan, D.C. 1999. Patterns of rational default. Regional Science and Urban Economics 29(6), 765-785.

Kau, J.B, Keenan, D.C., Muller, W.J., Epperson, J.F. 1992. A generalized valuation model for fixed-rate residential mortgages. Journal of Money, Credit and Banking 24(3), 279-299.

Keys, B., Mukherjee, T., Seru, A., Vig, V. 2010a. Did securitization lead to lax screening? evidence from subprime loans. Quarterly Journal of Economics 125, 307-362.

Keys, B., Mukherjee, T., Seru, A., Vig, V. 2010b. 620 FICO, take II: securitization and screening in the subprime mortgage. Review of Financial Studies, forthcoming.

Kotsiantis, S. B. 2007. Supervised Machine Learning: A Review of Classification Techniques, Informatica (31): 249-268.

Kum, H., & Masterson, T. 2008. Statistical matching using propensity scores: Theory and application to the levy institute measure of economic well-being (No. 535). Working papers, The Levy Economics Institute.

Langley, Pat, Simon, Herbert A. 1995. Applications of Machine Learning and Rule Induction, Communications of the ACM, 38(11):54-64.

Lauria, M. 1998. A new model of neighborhood change: reconsidering the role of white flight. Housing Policy Debate, 9(2), 395-424.

Lauria, M. and V. Baxter. 1999. Residential Mortgage Foreclosure and Racial Transition in New Orleans. Urban Affairs Review, 34(6): 757-786.

Lin, Z., Rosenblatt, E., & Yao, V. W. 2009. Spillover effects of foreclosures on neighborhood property values. The Journal of Real Estate Finance and Economics, 38(4), 387-407.

Liu, B. 2011. Credit risks and default behavior of mortgagors.

Lee, S. H. 2011. The Impact of Mortgage Foreclosures on Existing Home Prices in Housing Boom and Bust Cycles: A Case Study of Phoenix, AZ (Doctoral dissertation, Texas A&M University).

Leece, D. 2008. Economics of the mortgage market: perspectives on household decision making. John Wiley & Sons.

Leonard, T., & Murdoch, J. C. 2009. The neighborhood effects of foreclosure. Journal of Geographical Systems, 11(4), 317-332.

Lu, Q. and Getoor, L. 2003. “Link-based Classification,” in (T. Fawcett and N. Mishra, eds.) Proceedings of the Twentieth International Conference on Machine Learning, 496-503.

Mayer, C., Hubbard, G.R. 2010. House prices, interest rates, and the mortgage market meltdown, Working paper. Columbia University.

Mayer, C., Pence, K., Sherlund, S.M. 2009. The rise in mortgage defaults. Journal of Economic Perspectives 23(1), 27-50.

Mayer, C., Morrison, E., Piskorski, T., Gupta, A. 2011. Mortgage modification and strategic default: evidence from a legal settlement with Countrywide. Working paper.

Méray, N., Reitsma, J. B., Ravelli, A. C., & Bonsel, G. J. 2007. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. Journal of clinical epidemiology, 60(9), 883-e1.

Mian, A., Sufi, A., Trebbi, F. 2010. The political economy of the U.S. mortgage default crisis, American Economic Review 100(5), 1967-1998.

Mian, A., Sufi, A., Trebbi, F. 2014 Foreclosures, Foreclosures, House Prices, and the Real Economy, Journal of Finance, forthcoming.

Mian, A., Sufi, A., Trebbi, F. 2012. The political economy of the subprime mortgage credit expansion. Working paper.

Michie, Donald, David J. Spiegelhalter, and Charles C. Taylor. 1994. Machine learning, neural and statistical classification. Working paper.

Mitchell, Tom M. 1997. Machine learning. Burr Ridge, IL: McGraw Hill 45.

Moreno, A. 1995. The Cost-Effectiveness of Mortgage Foreclosure Prevention. Minneapolis: Family Housing Fund.

Murphy, K. P. 2012. Machine learning: a probabilistic perspective. The MIT Press.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, A. P. James. 1959. Automatic Linkage of Vital Records, Science, 130(3381):954-959.

Nilsson, N.J. 1965. Learning machines. New York: McGraw-Hill.

Pace, R. Kelley, Zhu, Shuang. 2012. Mortgage Choices and House Sales, Working Paper.

Patridge, C. 1998. The Fuzzy Feeling SAS Provides: Electronic Matching of Records without Common Keys. Working paper.

Piskorski, T., Seru, A., Vig, V. 2010. Securitization and distressed loan renegotiation: evidence from the subprime mortgage crisis. Journal of Financial Economics 97(3), 369-397.

Piskorski, T., Tchistyi, A. 2011. Stochastic house appreciation and optimal mortgage lending. Review of Financial Studies 24, 1407-1446.

Pollack, C. E., & Lynch, J. 2009. Health status of people undergoing foreclosure in the Philadelphia region. American Journal of Public Health, 99(10), 1833-1839.

Posner, R., Zingales, L. 2009. A loan modification approach to the housing crisis. American Law and Economics Review 11(2), 575-607.

Quercia, R. G., & Stegman, M. A. 1992. Residential mortgage default: a review of the literature. Journal of Housing Research, 3(2), 341-379.

Quigley, J. M., & Van Order, R. 1995. Explicit tests of contingent claims models of mortgage default. The Journal of Real Estate Finance and Economics, 11(2), 99-117.

Quinlan, J. R. 1986. Induction of decision trees, Machine learning: 1(1), 81-106.

Quinlan, J.R. 1993. C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.

Rajan, U., Seru, A., Vig, V. 2013. The Failure of Models That Predict Failure: Distance, Incentives and Defaults. Journal of Financial Economics, forthcoming.

Rässler, S. 2002. Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York: Springer.

RealtyTrac. 2009. Foreclosure Activity Increase 81 Percent in 2008 (2008 Metropolitan Foreclosure Market Report).

RealtyTrac. 2010. Year End Report Shows Record 2.8 Million U.S. Properties with Foreclosure Filings in 2009 (2009 Foreclosure Annual Market Report).

Riddiough, T. J., and S.B. Wyatt. 1994. Strategic Default, Workout, and Commercial Mortgage Valuation. Journal of Real Estate Finance and Economics, 9(1): 5-22.

Rogers, W. H., & Winter, W. 2009. The impact of foreclosures on neighboring housing sales. Journal of Real Estate Research, 31(4), 455-479.

Roos, LL; Wajda A. 1991. Record linkage strategies. Part I: Estimating information and evaluating approaches, Methods of Information in Medicine 30 (2): 117–123.

Rosenbaum, P. R., & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Rosenbaum, P. R. 2002. Observational Studies (2nd Ed.). New York: Springer-Verlag.

Rosenbaum, P. R. 2009. Design Observational Studies. New York: Springer-Verlag.

Rubin, D. B., & Thomas, N. 1992. Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika, 79(4), 797-809.

Rubin, D. B., & Thomas, N. 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics, 249-264.

Rubin, D. B. 2006. Matched Sampling for Causal Effects. Cambridge: Cambridge University Press.

Sanders, A. 2008. The Subprime Crisis and its Role in the Financial Crisis. Journal of Housing Economics, 17(4), 254-261.

Sarawagi, S. and Bhamidipaty, A. 2002. Interactive Deduplication Using Active Learning, Very Large Data Bases 2002, 269-278.

Schuetz, J., Been, V., and Ellen, I.G. 2008. Neighborhood Effects of Concentrated Mortgage Foreclosures. Journal of Housing Economics, 17(4): 306-319.

Simon, Phil. 2013. Too Big to Ignore: The Business Case for Big Data. Wiley.

Seiler, M. J., Collins, A. J., & Fefferman, N. H. 2013. Strategic Mortgage Default in the Context of a Social Network: An Epidemiological Approach. Journal of Real Estate Research, 35(4), 445-475.

Seiler, M. J., Lane, M. A., & Harrison, D. M. 2012. Mimetic herding behavior and the decision to strategically default. The Journal of Real Estate Finance and Economics, 1-33.

Singh, Yogesh, Pradeep Kumar Bhatia, and Omprakash Sangwan. 2007. A review of studies on machine learning Techniques. International Journal of Computer Science and Security 1.1: 70-84.

Shiller, R. J. 1995. Conversation, information, and herd behavior. The American Economic Review, 181-185.

Shiller, R. J. 2008. How a bubble stayed under the radar. The New York Times, 2.

Silveira, D. P. D., & Artmann, E. 2009. Accuracy of probabilistic record linkage applied to health databases: systematic review. Revista de Saúde Pública, 43(5), 875-882.

Steiner, Peter M., and David Cook. 2013. Matching and propensity scores. The Oxford.

Steiner, Peter M. 2011. Propensity Score Methods for Causal Inference: On the Relative Importance of Covariate Selection, Reliable Measurement, and Choice of Propensity Score Technique, Almalaurea Working Paper, University of Wisconsin–Madison.

Steyer, R. 2005. Analyzing individual and average causal effects via structural equation models. Methodology, 1, 39-64.

Taskar, B., Wong, M. F., and Koller, D. 2003. Learning on Test Data: Leveraging “Unseen” Features, Proceedings of the Twentieth International Conference on Machine Learning, 744-751.

Towe, C., & Lawley, C. 2013. The Contagion Effect of Neighboring Foreclosures. American Economic Journal: Economic Policy, 5(2), 313-335.

Vandell, K. D. 1993. Handing over the keys: a perspective on mortgage default research. Real Estate Economics, 21(3), 211-246.

Vandell, Kerry D. 1995. How Ruthless Is Mortgage Default? A Review and Synthesis of the Evidence. Journal of Housing Research 6(2): 245-264.

Van Rijsbergen, C. J., D. J. Harper, and M. F. Porter. 1981. The Selection of Good Search Terms,Information Processing and Management, 17, pp. 77-91.

Varian, H. 2005. Bootstrap Tutorial, Mathematica Journal, 9, 768-775.

Voicu, Ioan, Jacob, Marilyn, Rengert, Kristopher, Fang, Irene. 2011. Subprime Loan Default Resolutions: Do They Vary Across Mortgage Products and Borrower Demographic Groups? The Journal of Real Estate Finance and Economics.

Westreich, D., Lessler, J., & Funk, M. J. 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. Journal of clinical epidemiology, 63(8), 826.

White, B. T. 2010. Underwater and not walking away: shame, fear, and the social management of the housing crisis. Wake Forest L. Rev., 45, 971.

Winkler, William E. 1995. Matching and Record Linkage, U.S. Bureau of the Census.

Winkler, William. E. 2002. Record Linkage and Bayesian Networks, Proceedings of the Section on Survey Research Methods, American Statistical Association.

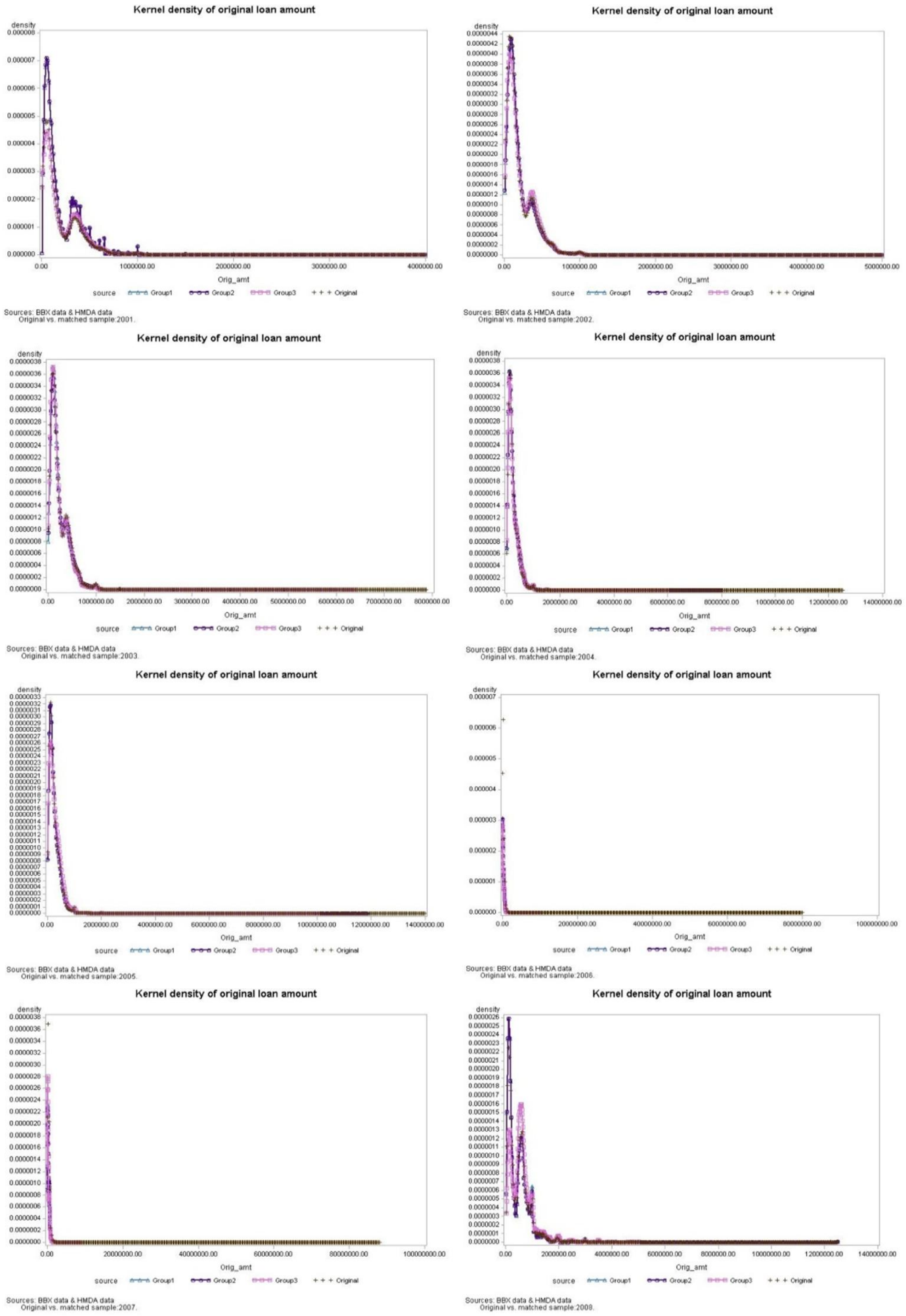
Winkler, William E. 2006. Overview of Record Linkage and Current Research Directions, Bureau of the Census.

Yu, C. T., Lam, K., & Salton, G. 1982. Term weighting in information retrieval using the term precision model. Journal of the ACM (JACM), 29(1), 152-170.

Zadrozny, B. 2004. Learning and Evaluating Classifiers under Sample Selection Bias, Proceedings of the International Conference on Machine Learning.

Zhou, Z., & Lam, P. 2007. Discussion of: Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. Journal of Biopharmaceutical Statistics, 17(1), 25-27.

Appendices



Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2001.

Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2002.

Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2003.

Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2004.

Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2005.

Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2006.

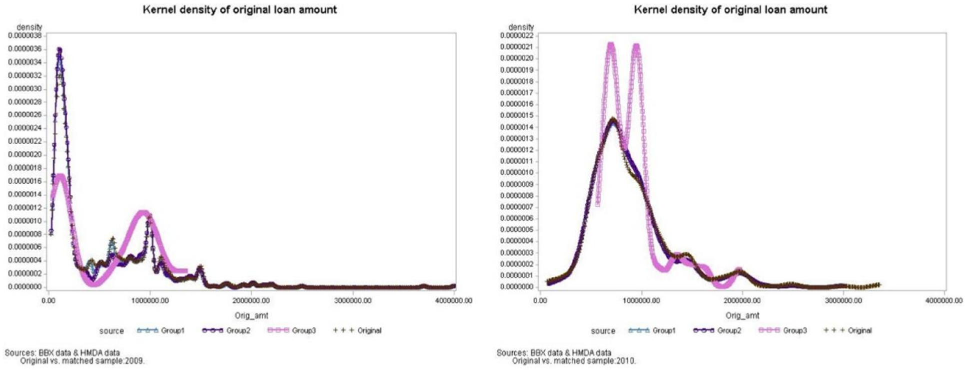
Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2007.

Kernel density of original loan amount

Sources: BBX data & HMDA data
Original vs. matched sample:2008.

Figure A1 Kernel Density Plots of the original loan amount from the Original and Matched Samples: by loan origination year (2001-2010)



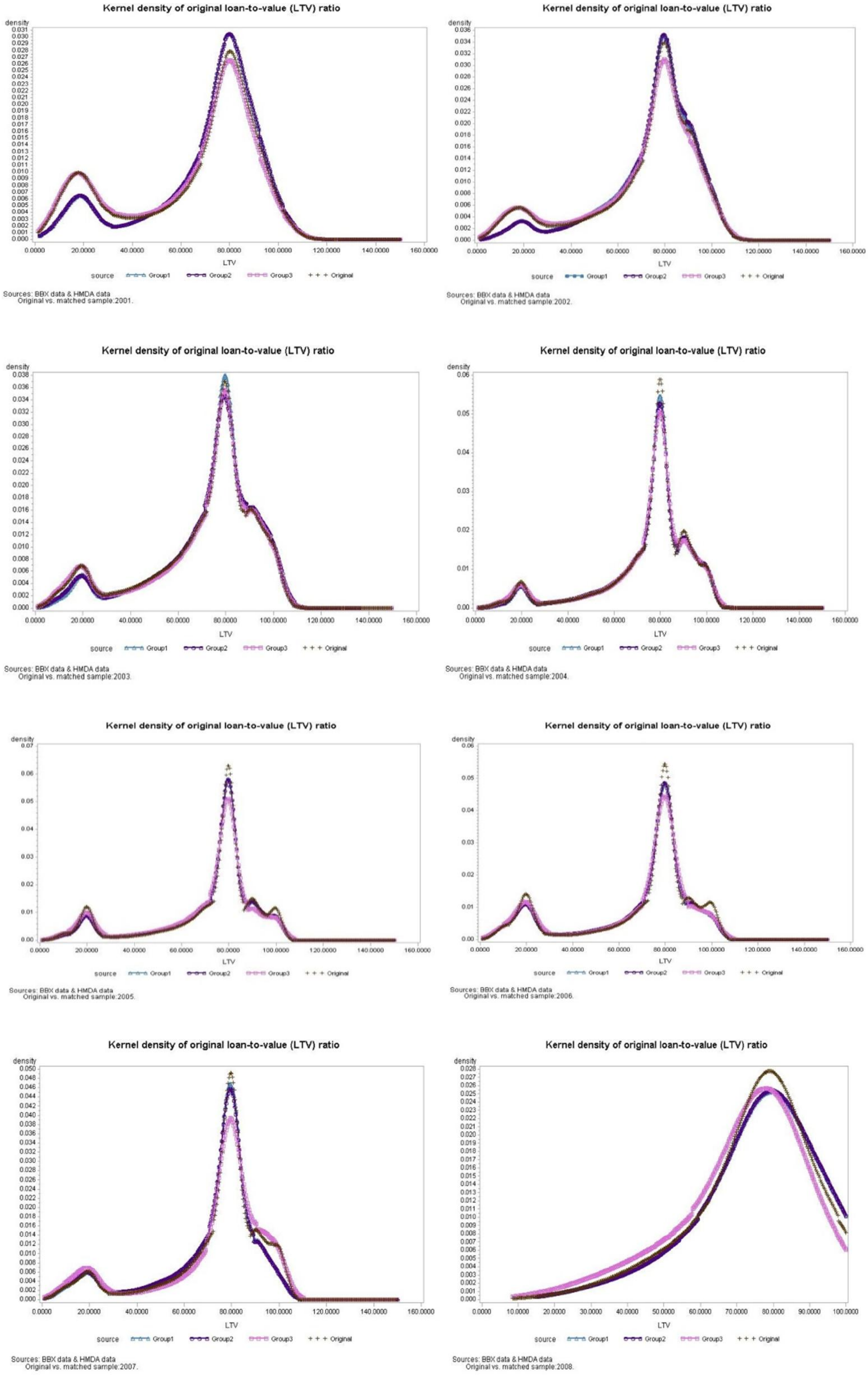
Sources: BBX data & HMDA data
Original vs. matched sample:2009.

Sources: BBX data & HMDA data
Original vs. matched sample:2010.

Figure A1 Kernel Density Plots of the original loan amount from the Original and Matched Samples: by loan origination year (2001-2010) (Continued)

Note:

This set of figures shows the kernel density plots of the key variable, original loan amount in BBX dataset, by origination (2001-2010). Four groups of data are compared: the original BBX dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Only loans with original loan amount less than $10 million are included in the sample. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.



Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2001.

Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2002.

Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2003.

Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2004.

Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2005.

Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2006.

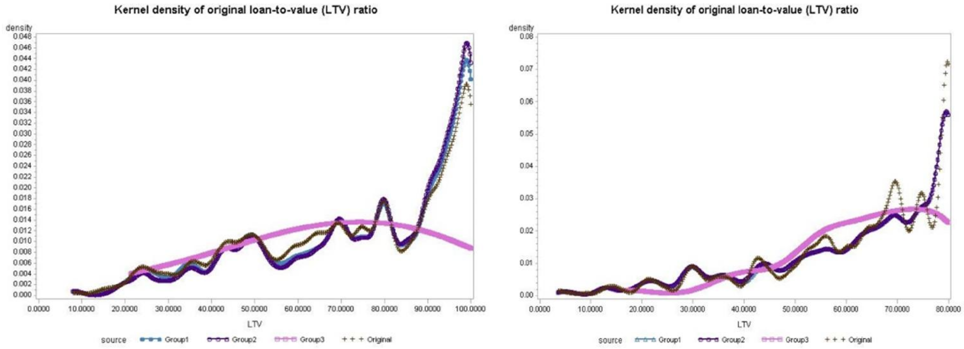
Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2007.

Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2008.

Figure A2 Kernel Density Plots of the original LTV ratio from the Original and Matched Samples: by loan origination year (2001-2010)



Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Original vs. matched sample:2009.

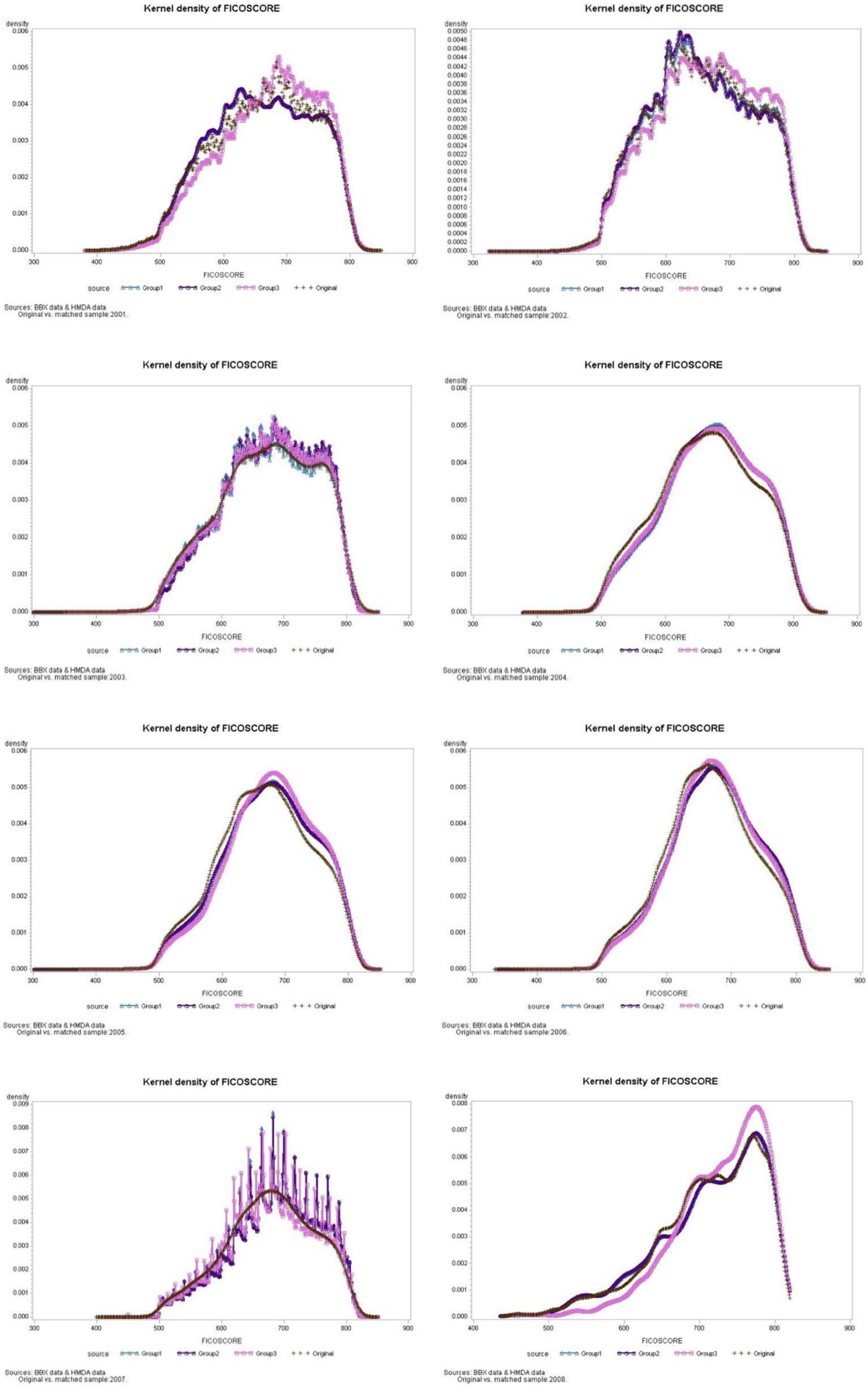
Kernel density of original loan-to-value (LTV) ratio

Sources: BBX data & HMDA data
Originalvs. matched sample:2010.

Figure A2 Kernel Density Plots of the original LTV ratio from the Original and Matched Samples: by loan origination year (2001-2010) (Continued)

Note:

This set of figures shows the kernel density plots of the key variable, original loan-to-value (LTV) ratio in BBX dataset, by origination (2001-2010). Four groups of data are compared: the original BBX dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Only loans with original loan amount less than $10 million are included in the sample. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.



Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2001.

Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2002.

Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2003.

Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2004.

Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2005.

Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2006.

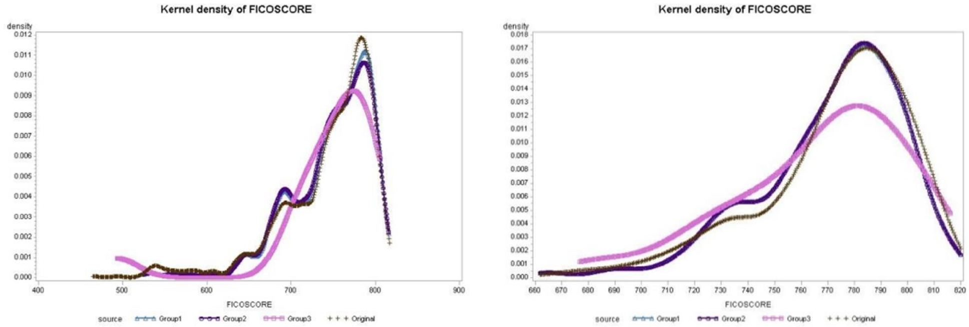
Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2007.

Kernel density of FICOSCORE

Sources: BBX data & HMDA data
Original vs. matched sample:2008.

Figure A3 Kernel Density Plots of the FICO score from the Original and Matched Samples: by loan origination year (2001-2010)



Sources: BBX data & HMDA data
Original vs. matched sample:2009.

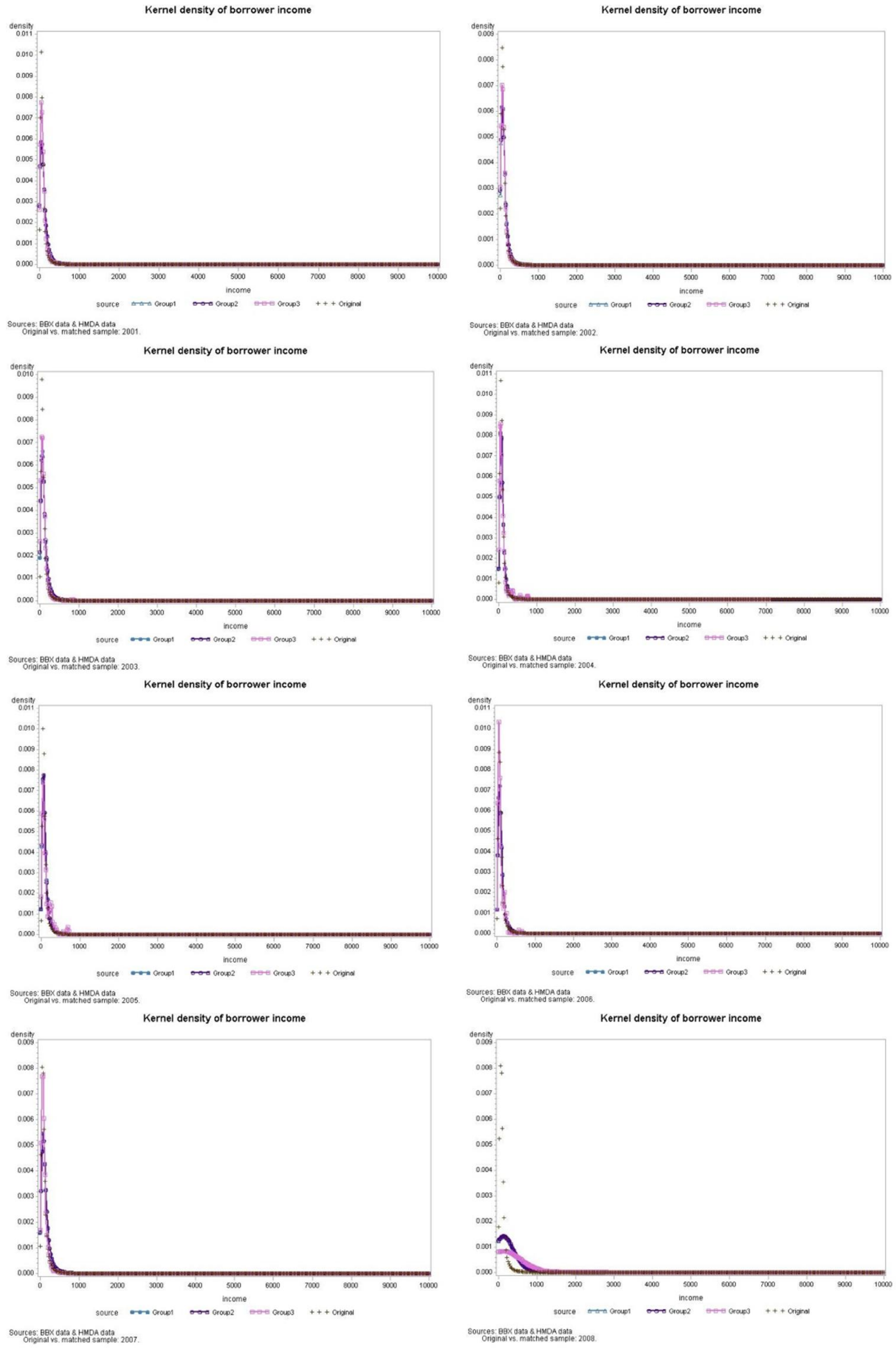
Sources: BBX data & HMDA data
Original vs. matched sample:2010.

Figure A3 Kernel Density Plots of the FICO score from the Original and Matched Samples: by loan origination year (2001-2010) (Continued)

Note:

This set of figures shows the kernel density plots of the key variable, original loan-to-value (LTV) ratio in BBX dataset, by origination (2001-2010). Four groups of data are compared: the original BBX dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Only loans with original loan amount less than $10 million are included in the sample. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.

Figure A4 Kernel Density Plots of borrower income from the Original and Matched Samples: by loan origination year (2001-2010)



Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2001.

Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2002.

Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2003.

Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2004.

Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2005.

Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2006.

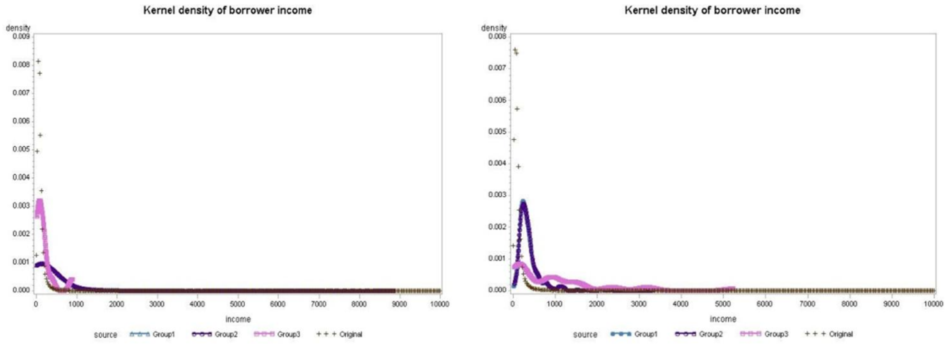
Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2007.

Kernel density of borrower income

Sources: BBX data & HMDA data
Original vs. matched sample: 2008.

Figure A4 Kernel Density Plots of borrower income from the Original and Matched Samples: by loan origination year (2001-2010)



Sources: BBX data & HMDA data
Original vs. matched sample: 2009.

Sources: BBX data & HMDA data
Original vs. matched sample: 2010.

Figure A4 Kernel Density Plots of borrower income from the Original and Matched Samples: by loan origination year (2001-2010) (Continued)

Note:

This set of figures shows the kernel density plots of the key variable, borrower income in HMDA dataset, by origination (2001-2010). Four groups of data are compared: the original HMDA dataset (Original), the matched sample from pure statistical hard matching (Group1), from machine learning techniques (Group2), and from propensity score matching (Group 3). Only loans with original loan amount less than $10 million are included in the sample. Y-axis indicates the probability of density, while X-axis indicates the value distribution of variables.

Table A1 Summary Statistics of Freddie Mac Full Sample

Panel A: Summary statistics for Freddie Mac: from 2003 to 2007				
	Total	Condo	Single-Family (SF)	Diff. (Condo-SF)
D_default within 2 yrs	1%	1%	1%	0%
FICO score	659	660	658	2***
Original LTV	75%	75%	75%	0***
Log_Original loan balance	11.91	11.87	11.91	-0.04***
Current interest rate	6.00	6.08	5.99	0.09***
D_Owner occupied	90%	78%	91%	-13%***
Log_HPI	5.19	5.24	5.18	0.06***
Log_duration	3.79	3.77	3.79	-0.02***
Sample Size (in thousands)	3,792	399	3,393	

Panel B: Summary statistics for Freddie Mac by loan origination year (2003–2007)

	Condo	SF	Diff.	Condo	SF	Diff.	Condo	SF	Diff.	Condo	SF	Diff.	Condo	SF	Diff.
Original Year	**2003**			**2004**			**2005**			**2006**			**2007**		
D_default within 2 yrs	0%	0%	0%***	0%	1%	0%***	0%	1%	0%***	1%	1%	0%***	3%	3%	0%**
FICO score	660	659	1***	659	658	1***	660	659	1***	660	659	1***	660	658	2***
Original LTV	74%	73%	1%***	75%	76%	-1%***	75%	75%	0%***	76%	76%	0%***	77%	77%	0%***
Log_Original loan balance	11.75	11.87	-0.12***	11.80	11.88	-0.08***	11.90	11.94	-0.04***	11.95	11.97	-0.02***	12.00	11.98	0.02***
Current interest rate	5.82	5.76	0.06***	5.89	5.85	0.04***	5.89	5.86	0.03***	6.46	6.44	0.02***	6.40	6.42	-0.02***
D_Owner occupied	82%	94%	-12%***	78%	91%	-13%***	77%	91%	-14%***	77%	90%	-13%***	76%	87%	-11%***
Log_HPI	5.27	5.19	0.08***	5.26	5.19	0.07***	5.22	5.17	0.05***	5.22	5.17	0.05***	5.20	5.15	0.05***
Log_duration	3.79	3.87	-0.08***	3.83	3.87	-0.04***	3.93	3.90	0.03***	3.72	3.64	0.08***	3.59	3.49	0.01***
Sample Size (in thousands）	106	1,153		67	628		75	630		75	494		76	488	

Note:

This table presents the summary statistics of the Freddie Mac sample. This dataset includes only single-family and condominium (condo) loans originated during the period 2003–2007. Panel A reports the results from aggregate-level summary statistics of the loans and compares the average values of the variables by full sample, single-family loans, and condo loans, respectively. Panel B shows the full sample summary statistics results by origination year. The variables with "D_" represent dummies. *D_default within 2 yrs* is equal to one for defaulting within two years of the loan origination date. *Current interest rate* refers to the coupon rate charged to the borrower for the most recent remittance period. *Log_Original loan balance* is defined as log of the amount of principal on the closing date of the mortgage. *FICO score* refers to the FICO (formerly the Fair Isaac Corporation) borrower credit score at the time of the loan closing. *Original LTV* means the ratio of the original loan amount to the property value at loan origination. *D_FRM* is equal to one for fixed-rate mortgages. *D_Owner occupied* takes one if the property is owner occupied. *Log_HPI* is log of the MSA-level quarterly FHFA/OFHEO House Price Index. *Log_Duration* is the log of the elapsed time from origination to the end of the sample period or to the first classification as being prepaid or delinquent at least 60 days. Note that ***, ** and * indicate 1%, 5% and 10% significance, respectively.