Journal of Financial Economics 115 (2015) 237–260

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

The failure of models that predict failure: Distance, incentives, and defaults ☆

Uday Rajan [a], Amit Seru [b,*], Vikrant Vig [c]

[a] Ross School of Business, University of Michigan, Ann Arbor, USA
[b] Booth School of Business, University of Chicago, Chicago, USA
[c] London Business School, London, UK

ARTICLE INFO

Article history:
Received 24 September 2011
Received in revised form
25 June 2012
Accepted 30 November 2012
Available online 19 September 2014

JEL classification:
G18
G21
G28

Keywords:
Statistical model
Lucas critique
Mortgage default
Regulation

ABSTRACT

Statistical default models, widely used to assess default risk, fail to account for a change in the relations between different variables resulting from an underlying change in agent behavior. We demonstrate this phenomenon using data on securitized subprime mortgages issued in the period 1997–2006. As the level of securitization increases, lenders have an incentive to originate loans that rate high based on characteristics that are reported to investors, even if other unreported variables imply a lower borrower quality. Consistent with this behavior, we find that over time lenders set interest rates only on the basis of variables that are reported to investors, ignoring other credit-relevant information. As a result, among borrowers with similar reported characteristics, over time the set that receives loans becomes worse along the unreported information dimension. This change in lender behavior alters the data generating process by transforming the mapping from observables to loan defaults. To illustrate this effect, we show that the interest rate on a loan becomes a worse predictor of default as securitization increases. Moreover, a statistical default model estimated in a low securitization period breaks down in a high securitization period in a systematic manner: it underpredicts defaults among borrowers for whom soft information is more valuable. Regulations that rely on such models to assess default risk could, therefore, be undermined by the actions of market participants.

1. Introduction

Statistical predictive models are extensively used in the marketplace by policy makers, regulators, and practitioners to infer the true quality of a loan. Such models are used by regulators to determine capital requirements for banks based on the riskiness of loans issued, rating agencies to predict default rates on underlying collateral, and banks

☆ For helpful comments and discussions, we thank numerous individuals, including two anonymous referees, as well as participants at seminars at Bank of London, University of California at Berkeley, Federal Reserve Board of Governors, Brigham Young University, Federal Reserve Bank of Chicago, Columbia University, Harvard University, Houston, London School of Economics, University of Michigan, Michigan State University, MIT Sloan School of Management, New York University Stern School of Business, University Naples Federico II, Federal Reserve Bank of Philadelphia, Stanford University, University of California at Los Angeles, University of Utah and at the American Economic Association, American Law and Economics Association, Rothschild Caesarea Center, European Finance Association, Financial Intermediation Research Society, Freiburg, Indian School of Business, London Business School and London School of Economics Credit Risk, National Bureau of Economic Research (NBER) Behavioral, NBER Summer Institute, Southwind Finance and Western Finance Association conferences. We are also indebted to Tanmoy Mukherjee for extensive discussions. All errors are our responsibility. Amit Seru thanks the Initiative on Global Markets at the University of Chicago for financial support. Vikrant Vig acknowledges the support provided by the Research and Materials Development grant at the London Business School. Part of this work was undertaken when Amit Seru was a visiting scholar at Sorin Capital Management.

* Corresponding author at: University of Chicago, 5807 S. Woodlawn Avenue, Chicago, IL 60637, USA.
E-mail addresses: urajan@umich.edu (U. Rajan), amit.seru@chicagobooth.edu (A. Seru), vvig@london.edu (V. Vig).

http://dx.doi.org/10.1016/j.jfineco.2014.09.012
0304-405X/

to decide what information they should collect to assess the creditworthiness of borrowers. In each case, the true quality of the loan might not be known for years, so participants in current transactions must rely on some observable features about the loan to assess the quality. For example, a bank regulator could consider the credit scores of borrowers and a collateralized debt obligation (CDO) investor could consider the interest rates on the underlying loans.

These statistical models have come under much scrutiny in the context of the subprime mortgage market, where they were extensively used to forecast the default likelihood of borrowers and of collateral. There has been a public outcry over the failure of rating agency models that estimate the quality of CDO tranches (see Faltin-Traeger, Johnson, and Mayer, 2010, and Griffin and Tang, 2012). In addition, statistical scoring models such as FICO credit scores that assess a subprime borrower's default probability and guide lender screening have come under scrutiny.[1] Why did statistical default models fare so poorly in the build-up to the subprime crisis? A common answer to this question is that they were undermined by unanticipated movements in the house prices (see, e.g., Brunnermeier, 2009). We argue that this is far from the complete story. Our central thesis is that a primary reason for the poor performance of these predictive models is that they are subject to the classic Lucas critique (Lucas, 1976): They fail to account for a change in the relations between variables when the behavior of agents that influence these relations changes.

We analyze this phenomenon in the context of subprime mortgage loans issued in the US over the period 1997–2006. A notable feature of this period is a progressive increase in the proportion of loans that are securitized. Securitization changes the nature of lending from "originate and hold" to "originate and distribute," and it increases the distance between a homeowner and the ultimate investor. A loan sale to an investor results in information loss: some characteristics of the borrower that are potentially observable by the originating lender are not transmitted to the final investor.[2] Because the price paid by the investors depends only on verifiable information transmitted by the lender, this introduces a moral hazard problem: The lender originates loans that rate high based on the characteristics that affect its compensation, even if the unreported information implies a lower quality. The same tension exists in the multitasking framework of Holmström and Milgrom (1990): An agent compensated for specific tasks ignores other tasks that also affect the payoff of the principal.

In general, the quality of a mortgage loan is a function of both hard and soft information that the lender can obtain about the borrower (see Stein, 2002). Hard information, such as a borrower's FICO credit score, is easy to verify; conversely, soft information, such as the borrower's future job prospects, is costly to verify (see, e.g., Agarwal and Hauswald, 2010; Liberti and Mian, 2009 on the role of soft information in the context of business lending). In the absence of securitization, a lender internalizes the benefits and costs of acquiring both kinds of information and adequately invests in both tasks. With securitization, hard information is reported to investors; soft information, which is difficult to verify and transmit, remains unreported. Investors, therefore, rely only on hard information to judge the quality of loans. This eliminates the lender's incentives to produce soft information.[3] Consequently, after a securitization boom, among borrowers with similar hard information characteristics, over time the set that receives loans becomes worse along the soft information dimension. That is, securitization changes the incentives of lenders, and hence their behavior. The result is a change in the relation between the hard information variables (such as the FICO score) and the quality of the loan (such as the likelihood of default). This implies a breakdown in the quality of predictions from default models that use parameters estimated using data from the pre-boom period.

We provide evidence for our thesis by demonstrating three main effects of increasing securitization over time. First, due to the greater distance between originators and investors, the interest rate on new loans depends increasingly on hard information reported to the investor. Second, due to the loss of soft information, the interest rate on a loan becomes an increasingly poor predictor of the likelihood of default on a loan. Third, because the change in lender behavior modifies the relation between observed characteristics of loans and their quality, a statistical model fitted on past data underestimates defaults in a predictable manner—precisely for those borrowers on whom soft information not reported to investors is likely to be important.

Our first result is that the mapping between borrower and loan characteristics and the interest rate on a loan changes with the degree of securitization. In setting the interest rate on a loan, the lender ceases to use information that is not reported to the final investor. Using a large database on securitized subprime loans across different US lenders, we find that over time the interest rate on new loans relies increasingly on a small set of variables. Specifically, the R^2 of a regression of interest rates on borrower FICO credit scores and loan-to-value (LTV) ratios increases from 9% for loans issued in the period 1997–2000 to 46% for 2006 loans. Further confirmation comes from the dispersion of interest rates: Conditioning on the FICO score, the standard deviation of interest rates on new loans shrinks over time. Finally, using data from a single large subprime lender, we demonstrate the converse: As securitization increases, interest rates depend less on information observed by the lender but unreported to investors.

[1] Calomiris (2009), Mayer (2010), and Pagano and Volpin (2010) discuss various issues and remedies related to the rating process.

[2] Bolton and Faure-Grimaud (2010) and Tirole (2009) argue that contracts will be endogenously incomplete when there are costs involved in verifying or processing information. Along similar lines, Stein (2002) draws a distinction between hard (verifiable) and soft (unverifiable) information. One can think of the latter as being verifiable only at an infinite cost; it cannot be communicated to a third party, and so cannot be contracted on.

[3] In the context of jumbo mortgage loans, Loutskina and Strahan (2011) suggest that geographic diversification adversely affects the ability to collect information about borrowers.

Second, we show that with increased securitization the interest rate becomes a worse predictor of default likelihood on a loan. With securitization, there is an information loss, because the lender offers the same interest rate to both good and bad types of borrowers (see Rajan, Seru, and Vig, 2010). As a result, in a high securitization regime, the interest rate becomes a noisier predictor of default for the loan pool. To demonstrate this, we regress actual loan defaults on the interest rate for loans in our main sample, where default is a binary variable considered in a two-year window from the issue date. We find that the pseudo-R^2 of this logit regression declines with securitization, confirming that the interest rate loses some of its ability to predict loan defaults.

Third, we show that the change in lender behavior as securitization increases alters the data generating process by transforming the mapping from all observables to loan defaults. We expect that reliance on past data will lead to underprediction of defaults in a high securitization regime, with the underprediction being more severe on borrowers for whom the unreported (or lost) information is more important. These borrowers include those with low FICO scores and high LTV ratios. To illustrate this effect, we estimate a baseline statistical model of default for loans issued in a period with a low degree of securitization (1997–2000), using information reported by the lender to the investor. We show that the model underpredicts defaults on loans issued in a regime with high securitization (2001 onward). The degree of underprediction is progressively more severe as securitization increases, indicating that, for the same observables, the set of borrowers receiving loans worsens over time. Further, we find a systematic variation in the prediction errors, which increase as the borrower's FICO score falls and the LTV ratio increases. As a placebo test, we estimate a default model for low-documentation loans over a subset of the low securitization era, and examine its out-of-sample predictions on loans issued in 1999 and 2000 (also a low securitization period). The statistical model performs significantly better than in our main test, and in particular yields prediction errors that are approximately zero on average.

We perform several cross-sectional tests to confirm our results. First, as a direct test of our information channel, we separately consider loans with full documentation and loans with low documentation. More information about a borrower is reported to investors on a full-documentation loan, including information on the borrower's income and assets. As a result, we expect that the prediction errors from the default model in the high securitization era should be lower for such loans. This is borne out in the data. Accounting for observables, the prediction errors on low-documentation loans are almost twice those on full-documentation loans during the high securitization regime.

Second, we perform two tests to rule out the concern that our findings on the performance of a statistical default model could be influenced by other macro factors that have changed over time with securitization. In the first test we compare loans securitized in states with foreclosure procedures that are more friendly to lenders with those issued in states with less lender-friendly procedures. Following Pence (2006) and Mian, Sufi, and Trebbi (2011), we compare loans in zip codes that border states with different foreclosure laws to account for both observable and unobservable differences across states. We postulate that lender-friendly foreclosures facilitate the securitization of loans, and we empirically confirm that the number of securitized loans (scaled by households) increases in lender-friendly states over time. Therefore, our expectation is that a statistical default model fitted to historical data should suffer a larger breakdown for loans in such states. This is confirmed by the data. The prediction errors from the default model are greater for loans made in lender-friendly states. Our second test has a similar flavor. We compare low-documentation loans whose borrowers have FICO scores just above 620 (which are easier to securitize; see Keys, Mukherjee, Seru, and Vig, 2010; Keys, Seru, and Vig, 2012) with those whose borrowers have FICO scores just below 620 (which are more difficult to securitize). We find that default prediction errors are higher for loans that are easier to securitize. Overall, these cross-sectional tests strongly corroborate our earlier findings.

Our baseline default model does not include the effects of changes in house prices, so one concern could be that a fall in house prices could lead to high defaults and explain most of the prediction errors in our analysis. It is important to note that several of our empirical strategies suggest otherwise. First, our cross-sectional tests compare loans in the same time period and with similar exposure to house prices. In addition, in the time series, we find that the default model underpredicts errors even in a period in which house prices were increasing (i.e., for loans issued in 2001–2004). Nevertheless, we also consider a stringent specification that both estimates the baseline model over a rolling window and explicitly accounts for the effects of changing house prices. We determine the statewide change in house prices for two years after the loan has been issued and include it as an explanatory variable in the default model (i.e., we assume perfect foresight on the part of regulators estimating the default model). Approximately 50% of the prediction error survives the new specification, and the qualitative results remain: A default model estimated in a low securitization regime continues to systematically underpredict defaults in a high securitization regime.

As long as soft information cannot be contracted upon, a securitizing lender has no incentive to collect it. This statement remains true even if rational investors anticipate higher default rates going forward and price loans accordingly. If investors are boundedly rational and underestimate future defaults, the moral hazard problem with respect to soft information collection is exacerbated. We examine the subordination levels of AAA-rated CDO tranches backed by subprime mortgage loans, and find essentially no relation between the mean prediction errors on defaults and subordination levels. This finding is consistent with rating agencies either being unaware of or choosing to ignore the adverse effects of securitization on the quality of the loan pool over time.

Our work directly implies that regulations based on statistical models can be undermined by the actions of market participants. For instance, the Basel II guidelines assign risk to asset classes relying in part on probability of default models.[4] We highlight the role of incentives in determining the riskiness of loans and, in turn, affecting the performance of models used to determine capital requirements. Our findings suggest that a blind reliance on statistical default models results in a failure to assess and regulate risks taken by financial institutions. Indeed, the regulation itself must be flexible enough for regulators to be able to adapt it to changing market circumstances (see Brunnermeier, Crockett, Goddhart, Persaud, and Shin, 2009 for another argument for flexible regulation).

More broadly, we identify a dimension of model risk (i.e., the risk of having an incorrect model) that cannot be corrected by mere application of statistical technique. The term "model risk" is often understood to refer to an incomplete set of data or conceptual errors in a model, or both. The focus in the literature has thus been on testing the consistency and robustness of inputs that go into statistical models. Collecting more historical data, possibly on extreme (and rare) events, is a key corrections that is frequently suggested. However, when incentive effects lead to a change in the underlying regime, the coefficients from a statistical model estimated on past data have no validity going forward, regardless of how sophisticated the model is or how well it fits the prior data. Indeed, aggregating data from different regimes may exacerbate the problem.

Although a naïve regulator might not understand that the lending regime has changed, we expect that rational investors will price loans accurately in either regime. Our hypotheses do not depend in any way on investors being boundedly rational.[5] However, if investors too are naïve, prices of loans or CDO tranches will fail to suitably reflect the default risk in a given loan pool. If anything, this exacerbates the tendency of lenders to stop screening borrowers on unreported information, leading to even greater underprediction of defaults. Misestimation of default risk by either regulators or investors could, in turn, lead to a misallocation of capital and a loss of welfare.

2. Hypothesis development

We start by examining how securitization changes the decision-making process of an originating lender, and thus affects the manner in which the interest rate evolves in our data. A lender has an imperfect screening technology that can generate two sets of observables, X_{it} and Z_{it}, on loan application i at time t. Here, observation i is a borrower–property pair; that is, the lender can acquire information both about a borrower and the property. Securitization entails the sale of the loan to an outside investor. If the loan is sold, the variables X_{it} are reported to the investor (so X_{it} must consist only of hard information), but the variables Z_{it} are not. Z_{it} could include both information variables that are quantified and maintained in the lender's own files (so are potentially verifiable by a third party) and soft information variables that are observed by neither the investor nor the econometrician.

On each loan application, the lender has two decisions to make: whether to approve the application and, if it does extend a loan, what interest rate to charge. Let A_{it} be a binary variable set to one if the application is approved and zero otherwise, and let r_{it} denote the interest rate on the loan. A lender's incentives to acquire and use information not reported to investors depend on the ease with which it securitizes loans on average.[6] As Keys, Seru, and Vig (2012) show, the ease of securitization can have multiple dimensions, including the probability or likelihood that a loan issued by a lender will be securitized and the average time taken to sell a loan. In this paper, for brevity we use the terms "high level of securitization" or "high securitization regime" to more generally mean a greater ease of securitization along all dimensions.

Intuitively, in a low securitization regime, both the approval decision and the interest rate depend on the variables X_{it} and Z_{it}. That is, we can write

$$A_{it} = f(X_{it}, Z_{it}), \tag{1}$$

$$\text{and} \quad r_{it} = g(X_{it}, Z_{it}). \tag{2}$$

As the level of securitization increases, a lender transits from a regime in which it retains most of the loans it issues to one in which it sells most of its loans. As it is costly to acquire information and the lender's own compensation on sold loans does not depend on the unreported variables Z_{it}, in a high securitization regime the lender stops collecting these variables. Its decisions now depend only on X_{it}, the variables that are reported to the investor. That is,

$$A_{it} = \tilde{f}(X_{it}), \tag{3}$$

$$\text{and} \quad r_{it} = \tilde{g}(X_{it}), \tag{4}$$

where we use the notation $\tilde{f}$ and $\tilde{g}$ to indicate that the mapping from the reported variables X_{it} to both the approval decision and the interest rate has changed after securitization.

Our first prediction is that, with increasing securitization, a focus on the variables X_{it} reported to the investor will lead to the offered interest rate relying to a greater extent on these variables. In a low securitization regime, if the interest rate is regressed only on the reported

[4] See, for example, Basel Committee on Banking Supervision (2006). Kashyap, Rajan, and Stein (2008) provide a detailed perspective on the role of capital requirements in the subprime crisis.

[5] While we are agnostic on whether investors mis-predicted the riskiness of loans in the build-up to the subprime crisis, emerging evidence shows that CDO tranches could have been mispriced (see, e.g., Benmelech and Dlugosz, 2009; Griffin and Tang, 2012; Faltin-Traeger, Johnson, and Mayer, 2010).

[6] We assume that, at the time a loan is issued, the lender does not know whether it will be securitized. In the subprime market, investors are typically offered a basket of loans and choose a subset of the basket. In addition, there is some quality checking through a comparison of loans sold by a lender and loans retained by it. It is difficult for lenders to cherry-pick loans to retain. This point is further discussed in Keys, Mukherjee, Seru, and Vig (2010) and Jiang, Nelson, and Vytlacil (2014).

variables, the estimated equation is $r_{it} = \hat{g}(X_{it})$. Because the interest rate also depends on the omitted variables Z_{it}, such a regression should provide a poor fit. In a high securitization regime, such a regression should yield a better fit, because the lender uses only X_{it} in setting the interest rate.

Our second prediction focuses on the relation between the interest rate and the probability that a loan will default. To understand the connection between the two we follow the theoretical underpinnings provided by Rajan, Seru, and Vig (2010). In particular, fix a value of X_{it}. For simplicity, assume that at that value of X_{it} there are two types of borrowers, with the good type always repaying the loan [and representing positive net present value (NPV) for a lender or investor] and the bad type always defaulting (so having a negative NPV). In a low securitization regime, the lender also acquires the information in Z_{it}, which provides a signal about type. Borrowers that generate a good signal are offered a low interest rate (say r_g) and those that generate a bad signal are screened out altogether.

In a high securitization regime, the lender no longer collects Z_{it}, so it must offer the same interest rates to both types. One possibility is to offer a high interest rate $\overline{r} > r_g$ that reflects the increased riskiness of the average borrower in the pool. However, borrowers with good types will refuse this offer—they are likely to obtain a loan at an interest rate r_g at some other lender, so their reservation rate is lower than that of bad types.[7] Then, if the pooled interest rate $\overline{r}$ is offered, only the bad types will accept and the lender will lose money. Instead, the lender must charge an interest rate that is sufficiently low to attract the good types as well. In particular, the lender must continue to offer the interest rate r_g.

Comparing across the low and high securitization regimes, therefore, defaults at the interest rate r_g will increase. In other words, the interest rate becomes a noisier predictor of defaults under high securitization and, in particular, underpredicts defaults.[8] We therefore predict that the relation between the interest rate and the actual default experience on loans becomes weaker as securitization increases.

Importantly, this intuition goes through even when investors have rational expectations and understand that the pool of borrowers has worsened in the high securitization regime. In a competitive market, investors are willing to buy the loan at a fair price—they are not fooled about the quality of the borrowers.

Our third prediction builds on the same intuition. Here, we focus on the mapping between all observables (including the interest rate) and loan defaults. We expect this mapping to change with securitization. To illustrate this in the data, we first fit a statistical default model to data generated in a low securitization regime and then consider the prediction errors the model generates on loans issued in a high securitization regime. As the interest rate does not change by enough to adequately reflect the worse quality of the loan pool, we expect the prediction errors (i.e., actual minus predicted defaults) to be positive on average. We also expect the prediction errors to increase with securitization and to be larger for borrowers on whom the unreported information is more informative about quality (in particular, borrowers with low FICO credit scores and high loan-to-value ratios).

In the Appendix, we explain how the change in the data generating process can be understood using the selection model framework of Heckman (1980). The essence of the argument is that a regulator and rating agencies see only approved loans, which are a selected sample. The approval process changes with lender incentives and behavior. Consequently, as securitization increases, one expects the change in lender behavior to affect the loans that are selected into the approved pool, thereby altering the mapping from observables to defaults.

3. Data

We use two sets of data in our analysis. Here, we describe the primary data set, which comes from LoanPerformance and is used in the bulk of the paper. A second data set consisting of loans from a single lender, New Century Financial Corporation (NCFC), is described in Section 4.3.

Our primary data set contains loan-level information on securitized non-agency mortgage loans. The data include information on issuers, broker dealers, deal underwriters, servicers, master servicers, bond and trust administrators, trustees, and other third parties. As of December 2006, there are more than eight thousand home equity and nonprime loan pools (more than seven thousand active) that include a total of 16.5 million loans (more than seven million active) with more than $1.6 trillion in outstanding balances. Estimates from LoanPerformance suggest that, as of 2006, the data cover over 90% of the subprime loans that have been securitized. As Mayer and Pence (2008) point out, there does not exist a universally accepted definition of "subprime." Broadly, a borrower is classified as subprime if she has had a recent negative credit event. Occasionally, a lender signals a borrower with a good credit score is subprime, by charging higher than usual fees on a loan. In our data, the vendor identifies loans as subprime or Alt-A (thought to be less risky than subprime, but riskier than agency loans).

The data set contains all variables obtained from the issuer by the investor, including the loan amount, maturity, loan-to-value (LTV) ratio, borrower credit score, interest rate, and other terms of the loan contract. The FICO credit score is a summary measure of the borrower's credit quality. This score is calculated using information about the borrower's credit history (such as the amounts of various types of debt outstanding), but not about her income or assets (see, for example, Fishelson-Holstein, 2005). The software used to generate the score from individual credit reports is licensed by the Fair Isaac Corporation to the three major credit repositories: TransUnion, Experian, and Equifax. FICO scores provide a ranking of potential borrowers by the probability of having any negative

[7] As discussed in Rajan, Seru, and Vig (2010), when it is costly for borrowers to search for loans, good types are likely to have lower search costs than bad types. As a result, they are able to obtain a lower interest rate at some other lender.

[8] Gorton and Pennacchi (1995) show that when screening is costly a lender exerts less effort on screening when it plans to sell a loan, so that the quality of the loan worsens. Along similar lines, Inderst and Ottaviani (2009) show that a lender who must compensate an agent for generating a loan reduces the standard of the issued loan.

credit event in the next two years. Probabilities are rescaled as whole numbers in a range of 400–900 (though nearly all scores in our data are between 500 and 800), with a higher score implying a lower probability of a negative event.

The LTV ratio of the loan, which measures the amount of the loan expressed as a percentage of the value of the home, also serves as a signal of borrower quality. For borrowers who do not obtain a second lien on the home, the LTV ratio provides a proxy for wealth. Those who choose low LTV loans are likely to have greater wealth and hence are less likely to default.

Borrower quality can also be gauged by the extent of documentation collected by the lender when approving the loan. The various levels are categorized as full, limited, or no documentation. Borrowers with full documentation provide verification of income as well as assets. Borrowers with limited documentation provide no information about income and some information about their assets. No-documentation borrowers provide no information about income or assets. In our analysis, we combine limited- and no-documentation loans and call them "low-documentation loans." Our results are unchanged if we remove the small proportion of loans that have no documentation.

Other variables include the type of the mortgage loan (fixed rate, adjustable rate, balloon, or hybrid) and whether the loan is provided for the purchase of a principal residence, to refinance an existing loan, or to buy an additional property. We present results exclusively on loans for first-time home purchases. We ignore loans on investment properties, which are more speculative in nature and likely to come from wealthier borrowers. The zip code of the property associated with each loan is included in the data set. Finally, there is information about the property being financed by the borrower and the purpose of the loan. As most loans in the data set are for owner-occupied single-family residences, townhouses, or condominiums, we restrict the loans in our sample to these groups. We also exclude non-conventional properties, such as those that are insured by the Federal Housing Administration or the Department of Veterans Affairs, pledged properties, and buy-down mortgages.

The data set has some limitations. We do not observe points or other up-front fees paid by borrowers. All variables pertaining to the loan and borrower are observed as of the time of loan origination, not kept track of dynamically over time. Thus, we do not observe changes in a borrower's FICO credit score after the loan has been issued. Finally, information on the cumulative loan-to-value (CLTV) ratio is not reliably present in the early part of the sample.

We consider only subprime mortgage loans in our analysis. We report year-by-year summary statistics on FICO scores and LTV ratios in Table 1. The number of securitized subprime loans increases more than fourfold from 2001 to 2006. This pattern is similar to that described by Demyanyk and Van Hemert (2011) and Gramlich (2007). The market has also witnessed an increase in the proportion of loans low (i.e., limited or no) documentation, from about 25% in 1997 to about 45% in 2006.

LTV ratios have gone up over time, as borrowers have put less equity into their homes at the initial purchase. The average FICO score of individuals who access the subprime market has been increasing over time, from 611 in 1997 to 636 in 2006. This increase in the average FICO score is consistent with a rule-of-thumb leading to a larger expansion of the market above the 620 threshold as shown in Keys, Mukherjee, Seru, and Vig (2010) and Keys, Seru, and Vig (2012). Though not reported in the table, average LTV ratios are lower and FICO scores higher for low-documentation loans, as compared with the full-documentation sample. This possibly reflects the additional uncertainty lenders have about the quality of low-documentation borrowers. The trends for loan-to-value ratios and FICO scores in the two documentation groups are similar.

Table 1
Summary statistics, primary data set.
This table reports summary statistics of FICO scores, loan-to-value (LTV) ratios and information on the documentation reported by the borrower (full, limited, or no) when taking the loan. Full-documentation loans provide verification of income as well as assets of the borrower. Limited documentation provides no information about the income but does provide some information about the assets. No documentation loans provide no information about income or assets. We combine limited and no documentation loans and call them "low-documentation loans."

Origination year	Number of loans	Proportion with low documentation (percent)	Mean LTV ratio (percent)	Mean FICO score
1997	24,067	24.9	80.5	611
1998	60,094	23.0	81.5	605
1999	104,847	19.2	82.2	610
2000	116,778	23.5	82.3	603
2001	136,483	26.0	84.6	611
2002	162,501	32.8	85.6	624
2003	318,866	38.9	87.0	637
2004	610,753	40.8	86.6	639
2005	793,725	43.4	86.3	639
2006	614,820	44.0	87.0	636

In Table 2, we report the proportion of newly issued subprime mortgage loans that are securitized in each period. The second row shows the overall securitization rate in the market; the third row the securitization rate for New Century Financial Corporation (NCFC). As shown in the table, both the overall market and NCFC experience a steady increase in the securitization rate over time. The securitization is relatively stable in the period 1997–2000, at around 37%, climbing to 76% in 2004 and even higher in 2006.

Together, the spikes in both the overall volume of loans and the securitization rate indicate that in the aggregate subprime market securitization had become an increasingly important phenomenon over this period. A common explanation for these trends (see, for example, Greenspan, 2008) is a surge in investor demand for securitized loans. Due to an unprecedented budget surplus, the US Treasury engaged in a buyback program for 30-year bonds in 2000–2001, and ceased to issue new 30-year bonds between August 2001 and February 2006 (Norris, 2006). Coincidentally, a rapid increase in CDO volume occurred over this period, with a significant proportion containing subprime assets.[9]

[9] The volume of CDOs issued in 2006 reached $386 billion, with home equity loans (largely from the subprime sector) providing for 26% of the underlying assets (see "Factbox – CDOs: ABS and other sundry collateral," reuters.com, June 28, 2007).

Table 2
Securitization rate over time.

This table reports the securitization rate for the overall subprime mortgage market and for New Century Financial Corporation (NCFC). The yearly securitization proportion for the overall market is obtained from *Inside B&C Lending*, a publication that has extensive coverage of the subprime mortgage market. Data on NCFC securitization rates comes from the origination and servicing loan files that encompass all lending activities of NCFC from 1997 to 2007.

	Securitization rate (percent)						
Origination year	1997–2000	2001	2002	2003	2004	2005	2006
Overall market	37	58	62	66	76	79	85
NCFC loans	41	50	77	88	92	85	96

It is important to remember that lenders in this market are heterogeneous and include commercial banks, thrifts, independent mortgage companies, and bank subsidiaries (see, for example, Gramlich, 2007). We expect that different lenders would cross over from a low to a high degree of securitization at different times. In addition, new lenders could enter the market over time. In both cases, we expect a lender securitizing a large proportion of loans to rely primarily on the variables reported to investors when issuing a loan and setting the interest rate on it. In the time series for the aggregate loan market, such behavior implies that our three hypotheses will hold on the entire sample.

The bulk of our tests, therefore, compare outcomes across time and examine whether incremental effects of increased securitization can be observed in the aggregate data. We consider the period 1997–2000 to be a low securitization regime and the period 2001 onward to involve high securitization.[10] In what follows, we use the term "year-by-year regression" to refer to separate regressions for the combined period 1997–2000 and for each year from 2001 to 2006.

4. Evolution of interest rate process: increased reliance on reported information

Our first prediction is that under high securitization interest rates will depend to a greater extent on variables that are reported to the investor. To test this prediction, we examine the evolution of the interest rate process over time. In Section 4.1, we consider our main sample. First, we directly regress the interest rate on a loan on the LTV ratio and the FICO score of the borrower. We predict that the explanatory power of the right-hand-side variables (i.e., the R^2 of the regression) will increase over time. We then consider the converse: If interest rates depend more on reported information as securitization increases, they must depend less on unreported information. Thus, keeping fixed the level of the reported variables such as the FICO score and the LTV ratio, interest rates should exhibit less dispersion at higher levels of securitization. In Section 4.3, we use our secondary data set of NCFC loans to examine the relation between interest rates and an internal ratings variable that is not reported to investors. In each of our tests, we find strong support for our prediction.

4.1. Relation between interest rate and reported variables: all subprime securitized loans

A direct way to capture the importance of the reported variables on the lender's behavior is to consider the R^2 of a year-by-year regression of interest rates on new loans on key variables. An increase in the R^2 of the regression over time indicates an increased reliance on variables reported to the investor.

We estimate the following regression year-by-year as our base model:

$$r_i = \beta_0 + \beta_{FICO} \times FICO_i + \beta_{LTV} \times LTV_i + \epsilon_i. \quad (5)$$

Here, r_i is the interest rate on loan i, $FICO_i$ is the FICO credit score of the borrower, LTV_i is the LTV ratio on loan i, and ϵ_i is an error term.

We report β_{FICO}, β_{LTV}, and the R^2 of the regression in Table 3. Consistent with our first prediction, Column 5 shows a dramatic increase in the R^2 of this regression over the years. Starting from about 9% in 1997–2000, the R^2 increases to 46.7% by the end of the sample. As expected, β_{FICO} is consistently negative (higher FICO scores obtain lower interest rates), and β_{LTV} is consistently positive (higher LTV ratios result in higher interest rates). Because the variance of FICO and LTV observed in the sample varies across years, the coefficients across years are not readily comparable. We re-estimate the base model after standardizing the interest rate, FICO score, and LTV ratio. The coefficients in the standardized regression also increase in magnitude over time. The R^2 of the standardized regressions is, of course, exactly the same as the R^2 reported in Table 3.

We next add dummy variables for three important features of the loan contract as explanatory variables to the base model: whether the loan is an adjustable rate mortgage (ARMs generally have low initial teaser rates), whether the loan has low documentation (full-documentation loans have lower interest rates), and whether there is a prepayment penalty. The R^2 of the enhanced model is reported in the Column 6 of Table 3. The added dummy variables somewhat improve the R^2 of the regression, but clearly preserve the trend, with the R^2 increasing from 11.4% in 1997–2000 to 50.8% in 2006. Although not reported in the table, the coefficients on the FICO score and LTV ratio for the regressions in the last two columns of the table are similar to those of the base model.

One concern could be that the results in the base model are driven by a change in lender composition over time instead of a change in lender behavior. To alleviate this concern, we estimate the base model using a fixed set of lenders across the sample period. The sample has several thousand lenders, each identified by name.[11] Most lenders are small: The largest 102 lenders account for

[10] In the overall market, the securitization rate over the period 1997 to 2000 remains between 33% and 41%. Because the volume of loans in each year in this period is also lower than in the later years, we combine these years in the rest of our analysis.

[11] The process of matching lenders to loans is somewhat cumbersome, because the same lender is sometimes referred to by slightly different names. For example, New Century Financial Corporation is sometimes referred to as New Century, NCF, and NCFC.

Table 3
Reliance of interest rates on FICO scores and loan-to-value (LTV) ratios.
This table reports estimates from the yearly regression of interest rates on FICO and LTV, using our primary data set. Standard errors are in parentheses. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

Origination year	Base model coefficients		Number of observations	Adjusted R^2 (percent) of various models		
	β_{FICO}	β_{LTV}		Base model	With additional contract variables	Including only lenders making 80% of loans
1997–2000	−0.009***	0.033***	305,786	8.98	11.38	8.40
	(0.0001)	(0.0003)				
2001	−0.012***	0.038***	136,483	19.49	22.74	20.13
	(0.0001)	(0.0004)				
2002	−0.011***	0.071***	162,501	17.42	26.43	15.66
	(0.0001)	(0.0001)				
2003	−0.012***	0.079***	318,866	29.72	41.26	33.29
	(0.0001)	(0.0001)				
2004	−0.010***	0.097***	610,753	36.85	45.39	41.00
	(0.0001)	(0.0001)				
2005	−0.009***	0.110***	793,725	43.91	50.14	52.82
	(0.0001)	(0.0001)				
2006	−0.011***	0.115***	614,820	46.67	50.83	46.72
	(0.0001)	(0.0001)				

approximately 80% of the data; the largest seven hundred lenders, for approximately 90% of the data. We re-run the regression including only the lenders comprising 80% of the loans and report the results in the last column of Table 3. As seen from the table, the R^2 displays the same trend as in the base model, suggesting that underlying our results is a change in lender behavior.

To ensure that our results are not driven simply by a change in the composition of the pool of loans over time, we estimate Eq. (5) year-by-year only for fixed rate mortgages. The R^2 increases from 11.0% for 1997–2000 loans to 36.8% for 2003 loans, and remains around 36% thereafter. The trend in the R^2 of the regression is therefore similar to that reported in the last three columns of Table 3. We also estimate Eq. (5) separately for loans with low documentation and those with full documentation, and find similar results. For brevity, these results are not reported in detail.

Finally, to the extent that the interest rate spread is the direct compensation to the lender for bearing the risk of the loan, we consider Eq. (5) with the interest rate spread instead of the raw interest rate as the dependent variable. In this regression, we consider only fixed rate mortgages and define the spread to be the difference between the mortgage interest rate and the ten-year current coupon Treasury rate. The results are very similar to the regression on fixed rate mortgages with the raw interest rate as the dependent variable, with the R^2 increasing from 10.9% in 1997–2000 to 36.2% in 2003, and remaining around 36% thereafter.

Across the various specifications, we consistently find that, in the low securitization regime (1997–2000), the variables reported to the investor explain very little variation in interest rates. The clear suggestion is that the unreported variables are particularly important in these years. As the securitization regime shifts, the same reported variables explain a large amount of variation in interest rates. Our results are thus consistent with the notion that the importance of variables not reported to the investor in determining interest rates on new loans declines with securitization.

In related work, Loutskina and Strahan (2011) find that banks that concentrate lending in a small number of markets are better able to price jumbo mortgage loans, which are more sensitive to soft information. In a different context, Cole, Goldberg, and White (2004) and Liberti and Mian (2009) find that loan offers to firms by large banks and at higher levels within a bank are more sensitive to financial statement variables, consistent with the notion that soft information cannot be communicated up the hierarchy within a firm.

4.2. Shrinkage of the distribution of interest rates

Another way to test the relation between information reported to investors and interest rates is to consider the dispersion of interest rates at different values of a reported variable. We calculate the standard deviation of interest rates at each FICO score and track it over time. Let $\sigma_{it} = \sqrt{(1/N)\sum_{j=1}^{N}(r_{ijt}-\bar{r}_{it})^2}$, where r_{ijt} is the interest rate on the jth loan with FICO score i in year t and $\bar{r}_{ijt} = (1/N)\sum_{j=1}^{N} r_{ijt}$ is the mean interest rate. We pool observations into FICO score buckets of 30 points starting from a score of 500 and ending at 799 (i.e., the buckets are FICO scores 500–529, 530–559, and so on). We then estimate the following regression separately for each bucket b:

$$\sigma_{bt} = \alpha_b + \beta_b \times t + \epsilon_{bt}, \tag{6}$$

where t indexes year and ϵ_{bt} is an error term. The coefficient β_b captures how the dispersion of interest rates within each FICO score bucket changes over time. We expect β_b to be large and negative for low FICO scores, i.e., we expect a shrinkage of dispersion in interest rates at low FICO scores. Information not reported to investors is likely to be more important in assessing the quality of such borrowers, compared to those with high FICO scores. We have ten

Table 4
Shrinkage in the distribution of interest rates.

We report estimates from a regression of yearly standard deviation of interest rates at different FICO scores on time. The regressions are estimated separately in buckets of 30 FICO points, starting with the bucket 500–529 and ending with 770–799. We include loans originated between 1997 and 2006. Standard errors are in parentheses. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

FICO score bucket	β_b	Standard error	R^2 (percent)
500–529	−0.209***	0.044	70.7
530–559	−0.168***	0.027	81.1
560–589	−0.099***	0.026	59.3
590–619	−0.042	0.025	17.0
620–649	−0.028	0.019	12.4
650–679	−0.055**	0.021	38.7
680–709	−0.058***	0.020	45.9
710–739	−0.079***	0.025	49.0
740–769	−0.085***	0.029	45.9
770–799	−0.065***	0.013	73.2

observations (one for each year from 1997 through 2006) in each bucket when we estimate Eq. (6). Given the potential for lack of power, the results of this subsection should be interpreted with caution.

We report the β_b coefficient for each FICO bucket in Table 4. For loans at low FICO scores (500–559), we find β_b to be about −0.17 to −0.2 (which translates to about a 7–8% reduction per year in the dispersion of interest rates). For higher FICO scores (560 and above), β_b is between −0.02 and −0.10 (a 2–4% reduction per year in the dispersion of interest rates). The magnitude of shrinkage can also be interpreted relative to the mean interest rate. Across sample years, the mean interest rate is 9.2% at FICO scores 500–559 and 8.1% at FICO scores 560 and higher. Thus, scaling the degree of shrinkage by the mean interest rate yields the same results.

4.3. Relation between interest rates and unreported variables: evidence from New Century Financial Corporation

In our primary data set, we do not observe variables that are not reported to investors, so we cannot directly demonstrate that the reliance on these variables reduces over time. We now examine data from a single lender, New Century Financial Corporation, which both confirm and enhance our findings. NCFC was a large subprime mortgage lender that filed for bankruptcy in April 2007.[12]

The NCFC data have two distinctive features that allow us to test our first hypothesis more extensively. First, the data contain both accepted and rejected loan applications, as well as both securitized loans and loans retained by NCFC. This allows us to directly consider the accept or reject decision and also to compute the proportion of securitized loans in each year. Second, and more important, the data set includes several variables that are not passed to investors but are observed by NCFC. Most important of these is an internal rating measure, which is assigned directly by NCFC loan officers. We expect the rating to summarize all relevant information about the loan available to a loan officer. This information includes variables that were passed on to investors (such as the FICO score and the LTV ratio). The rating ranges between 1 (best quality loan) and 20 (worst quality loan). Importantly, the measure is correlated with numerous variables contained in the NCFC data set (and therefore observed by NCFC) that are not reported to investors, including whether the borrower is self-employed, is married, has been referred by an existing customer, and has other debt in addition to the mortgage. We expect the rating to also capture soft information observed by NCFC but unobservable to both investors and the econometrician (such as a loan officer's assessment of default likelihood based on a personal interview with the borrower).

In the second row of Table 2, we report the proportion of loans issued by NCFC each year that are securitized. The results are consistent with the trend in the overall market: The proportion of securitized loans increases from 41% in the period 1997–2000 to 92% in 2004 and 96% in 2006. The overall summary statistics for securitized loans issued by NCFC are also similar to those reported for the aggregate market in Table 1. For example, the mean FICO score is 611 in the period 1997–2000, and 636 in 2006. Similarly, the mean LTV ratio is 79% in 1997–2000 and 85% in 2006.

To examine whether NCFC increasingly relies on the variables reported to the investor (specifically, the FICO score and the LTV ratio) in setting the interest rate on new loans, we estimate our base model in Eq. (5) on first-lien loans in the NCFC data, applying the same filters as in the main sample. The results are shown in Panel A of Table 5. The increase in the R^2 of the regression, from 10.8% in 1997–2000 to 28.1% in 2004, has a similar pattern to that shown for the aggregate market in Table 3, though the magnitude of the increase is somewhat smaller.

We now conduct two tests which directly provide evidence that the internal rating, which encapsulates several of the variables not reported to investors, increasingly becomes less important in the decisions made by NCFC. In the last column of Panel A of Table 5, we show the R^2 of the regression when the rating is added as an explanatory variable. The improvement in R^2 over the base model is about 50% for the period 1997–2000 and falls to 5% or less in the years 2004 through 2006. The results are therefore strongly consistent with NCFC abandoning its internal rating measure in setting interest rates, and relying instead on the FICO score and the LTV ratio.[13]

Next, we estimate a logit regression of the accept or reject decision on the internal rating measure. The regression equation here is

$$\text{Prob}(Accept_{it}=1)=\Phi(\beta_0+\beta_{\text{Rating}}Rating_{it}), \tag{7}$$

where $Accept_{it}$ is a binary variable equal to one if loan application i at time t was accepted, and zero otherwise, $Rating_{it}$ is the internal rating of application i at time t, and $\Phi(\cdot)$ is the logistic distribution function. The results are

[12] In 2006, NCFC had the second-highest market share in the US subprime mortgage market. See, for example, "New Century, biggest subprime casualty, goes bankrupt," bloomberg.com, April 2, 2007.

[13] Although not reported in the table, the coefficients on FICO score and LTV ratio are similar to those in the base model.

Table 5
Reliance of interest rates on reported and unreported variables.

This table reports results from the New Century Financial Corporation sample. For NCFC, we have information on accepted and rejected loan applications and on variables reported to investors as well as variables that are collected by the lender but not reported to investors. Panel A shows the coefficients and adjusted R^2 from an OLS regression of interest rates on the FICO score and LTV ratio and (last column) the internal rating measure. Panel B shows the coefficients from a logistic regression of the accept or reject decision for a loan application on the internal rating measure. Standard errors are in parentheses. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

Panel A: OLS regression of interest rate on FICO score and LTV ratio

Origination year	Base model coefficients		Number of observations	Adjusted R^2 (percent)	
	β_{FICO}	β_{LTV}		Base model	Model including internal rating
1997–2000	−0.0053***	0.014***	21,553	10.8	16.3
	(0.0001)	(0.0008)			
2001	−0.0072***	0.013***	7,302	12.9	18.9
	(0.0002)	(0.0016)			
2002	−0.0084***	0.009***	15,092	19.5	24.5
	(0.0001)	(0.0010)			
2003	−0.0085***	0.020***	33,690	25.1	28.6
	(0.0001)	(0.0006)			
2004	−0.0075***	0.050***	63,174	28.1	29.3
	(0.0001)	(0.0005)			
2005	−0.0062***	0.060***	84,002	23.9	24.4
	(0.0001)	(0.0005)			
2006	−0.0064***	0.066***	82,163	27.4	28.0
	(0.0001)	(0.0005)			

Panel B: Logit regression of accept/reject decision on internal rating measure

Origination year	β_{Rating}	Number of observations	Pseudo-R^2 (percent)
1997–2000	−0.053***	60,049	1.00
	(0.002)		
2001	−0.059***	14,905	1.12
	(0.004)		
2002	−0.070***	29,656	1.08
	(0.003)		
2003	−0.097***	71,188	0.76
	(0.004)		
2004	−0.075***	154,893	0.21
	(0.004)		
2005	−0.080***	199,369	0.16
	(0.004)		
2006	−0.056***	210,856	0.09
	(0.004)		

reported in Panel B of Table 5. While the coefficient β_{Rating} remains statistically significant in each year of the sample, the pseudo-R^2 of the regression falls from 1% or higher in the period 1997 through 2002 to 0.2% in 2004 and 0.09% in 2006. Therefore, over time, the internal rating measure becomes less important in the selection process for new loans.[14] We also find a dramatic decline in the variance of the internal rating measure over time. The variance declines by a factor of 15 between 1997–2000 and 2006. This is consistent with all loans being internally rated as high, so that the measure is not useful in setting the interest rate on a loan.

One could conjecture that the patterns observed both in the main and the NCFC data merely reflect that the FICO score is becoming a better predictor of defaults over time. If that were correct, lenders would need to collect and use less additional information in later years. However, we should then find that the FICO score becomes a better predictor of contemporaneous defaults over time. We estimate a logit regression of loan default within 24 months of origination on the FICO score, and find the exact converse. The pseudo-R^2 of the regression progressively falls from about 5% in 1997–2000 to 0.01% in 2006 for the main sample. The trend is similar in the NCFC data. Thus, we find that over time the FICO score becomes a poorer, not a better, predictor of loan defaults.

5. Evolution of default process

We now consider the effect of securitization on mortgage defaults. Following the arguments in Section 2, we

[14] Consistent with our other results, the accept or reject decision increasingly relies on the FICO score and LTV ratio over time. In a similar vein, when we regress loan defaults on the internal rating measure, we find that the measure progressively becomes a noisier predictor of defaults.

have two predictions on the default rates of loans. First, the ability of the interest rate to predict defaults should fall over time as information not being reported to the investor is no longer collected by the lender. Thus, in a year-by-year regression of default rates on interest rates, the R^2 should decrease over time. To test this prediction, we directly consider the evolution of the default process over time, as a function of the interest rate alone.

Second, we predict that the mapping between defaults and all observables changes with securitization. In particular, the quality of the loan pool should worsen, keeping fixed the observable characteristics of a loan. To test this prediction, we estimate a baseline statistical model using observables from a low securitization regime. We expect this baseline model to underpredict defaults under high securitization for borrowers on whom information not reported to investors is likely to be important in assessing quality; i.e., borrowers with low FICO scores and high LTV ratios.

5.1. Ability of interest rates to predict defaults

We examine the default experience of loans by issue year, assigning a variable $Actual\ Default_{it} = 1$ if loan i issued in year t defaults within 24 months of issue and zero otherwise. Here, default is defined to be the event that the loan is delinquent for at least 90 days. FICO scores are designed to predict negative credit events over the next two years.[15] Further, 24 months is before the first reset date of the most common types of ARMs in this market. We therefore restrict attention to defaults that occur within 24 months of loan origination.

The actual default experience on a loan in the two years beyond issue depends on many factors, including local and macroeconomic conditions and idiosyncratic shocks to the borrower's financial status. At the time the loan is issued, the interest rate on the loan reflects the lender's estimate of the overall likelihood the loan will default at some later point. It captures both what the lender knows about the riskiness of the borrower and the lender's forecast about future economic conditions that could influence default. Thus, we expect that the interest rate on a loan will be the most important predictor of whether the loan defaults.

Our hypothesis is that the interest rate loses its ability to predict defaults over time. We expect the loss of predictive ability to be more pronounced when the information not reported to the investor is more economically relevant, that is, for low-documentation loans and loans to borrowers at the lower part of the credit distribution. We therefore consider low- and full-documentation loans separately in our test and focus on the change in sensitivity of defaults to interest rates for borrowers at the 25th percentile of the FICO score distribution.

We estimate the following year-by-year logit model:

$$\text{Prob}(\text{Actual Default}_{it} = 1) = \Phi(\beta_0 + \beta_r r_{it}), \tag{8}$$

where r_{it} is the interest rate on loan i issued at time t.

[15] Holloway, MacDonald, and Straka (1993) show that the ability of FICO scores observed at loan origination to predict mortgage defaults falls by about 25% once one moves to a three-to-five year performance window.

Table 6
Contemporaneous default regressions.
This table reports the coefficients and pseudo-R^2 from a logistic regression of actual defaults on loan interest rates. A loan is defined to be in default if it is delinquent for at least 90 days within 24 months from the year of origination. Standard errors are in parentheses. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

Panel A: Low-documentation loans

Origination year	β_r	Constant (β_0)	Pseudo-R^2 (percent)	Number of observations
1997–2000	0.282***	−4.996***	2.43	65,895
	(0.00920)	(0.0965)		
2001	0.333***	−5.159***	3.42	35,110
	(0.0112)	(0.113)		
2002	0.224***	−4.079***	2.54	52,967
	(0.00709)	(0.0689)		
2003	0.224***	−4.023***	2.21	123,766
	(0.00514)	(0.0442)		
2004	0.159***	−3.215***	1.12	248,839
	(0.00341)	(0.0282)		
2005	0.127***	−2.331***	0.73	343,581
	(0.00247)	(0.0208)		
2006	0.111***	−1.444***	0.65	270,284
	(0.00231)	(0.0215)		

Panel B: Full-documentation loans

Origination year	β_r	Constant (β_0)	Pseudo-R^2 (percent)	Number of observations
1997–2000	0.211***	−4.065***	1.94	231,103
	(0.00376)	(0.0409)		
2001	0.243***	−4.051***	2.61	98,751
	(0.00506)	(0.0534)		
2002	0.177***	−3.344***	1.88	107,648
	(0.00437)	(0.0422)		
2003	0.240***	−3.856***	2.93	194,010
	(0.00355)	(0.0307)		
2004	0.199***	−3.268***	1.83	360,646
	(0.00261)	(0.0212)		
2005	0.140***	−2.451***	0.92	448,422
	(0.00215)	(0.0177)		
2006	0.0858***	−1.689***	0.38	343,393
	(0.00216)	(0.0199)		

Table 6 shows the estimated coefficients and the pseudo-R^2 values. First, consider Panel A, which reports on low-documentation loans. The pseudo-R^2 consistently falls from 3.42% for 2001 vintage loans to 1.12% for 2004 vintage loans and 0.65 for 2006 vintage loans. Further, at the 25th percentile of the FICO score distribution, a 1 standard deviation change in interest rate implies a change in default rate of about 4.2% in 2001, 2.0% in 2004, and 1.7% in 2006. That is, there is a decline in the sensitivity of defaults to interest rates in the later years of the sample, suggesting that interest rates are not responding as much to changes in the riskiness of a borrower. Defaults on loans issued in 2005 and 2006 are high from July 2007 onward in part due to a downturn in house prices. Although these two years are arguably special, it is important to note that the trends in both R^2 and the marginal effects of the coefficients are observable even over the period 2001–2004.

The results on full-documentation loans are shown in Panel B of Table 6. Among loans of vintage 2001 through 2004, no monotone pattern emerges in the R^2 of the

regression. Loans issued in 2005 and 2006 display the same trend as exhibited by low-documentation loans. Importantly, the marginal effect of the coefficients evaluated at the lower part of the credit distribution again suggests a progressive reduction in the sensitivity of interest rates to default risk. At the 25th percentile of the FICO score, the marginal effect of a 1 standard deviation change in the interest rate on the default rate is about 3.8% in 2001, 2.7% in 2004 and 1.9% in 2006.

5.2. Failure to predict failure: main test

We now test whether the mapping between observables reported to the investors and loan defaults has changed, by evaluating how a statistical default model estimated on historical data from a low securitization regime performs as securitization increases. In particular, we examine if the statistical model produces positive errors on average and whether these errors exhibit the systematic variation with observables predicted by our hypothesis. The exact nature of the statistical model used to assess our prediction is not important. The changed mapping between observables and defaults should show up in any statistical model that does not account for the effect of the increased distance between the borrower and the final investor on the incentives of the originating lender.

We consider the period 1997–2000 to be a low securitization era, and the period 2001–2006 to be a high securitization one. We estimate the following logit model on all securitized loans in our primary data set issued in the period 1997 to 2000:

$$\text{Prob}(\text{Actual Default}_i = 1) = \Phi(\beta \cdot X_i + \beta^{Low} \cdot I_i^{Low} X_i). \quad (9)$$

Here, X_i is a vector that includes the interest rate on the loan, the FICO credit score of the borrower, the LTV ratio, an ARM dummy, and a prepayment penalty dummy. I_i^{Low} is a dummy set to one if loan i has low documentation and zero otherwise. We also include state fixed effects in the regression. This model resembles the LEVELS® 6.1 Model used by Standard & Poor's. What is important here is not the exact specification of the model, but its use of historical information without regard to the changing incentives of agents who produce the data. The latter feature is common to most models used by rating agencies or regulators.

Panel A of Table 7 shows the estimated coefficients on the interest rate, FICO score, and LTV ratio from the baseline model. A low interest rate and high credit score are both associated with lowering the probability that the borrower will default in the subsequent two years, for both full-documentation and low-documentation loans.

Next, we use the coefficients of the baseline model to predict the probability of default for loans issued from 2001 to 2006, where default again is an event that occurs up to two years after a loan is issued. Concretely, let $\hat{\beta}_{1,t}$ and $\hat{\beta}^{Low}_{1,t}$ be the coefficients estimated from Eq. (9) for the baseline model over the period 1 to t (where year 1 is 1997 and year t is 2000). Then, for $k = 1, 2, \ldots, 6$, we estimate the predicted probability that a loan i issued at $t+k$ will default in the next 24 months (keeping the baseline coefficients fixed) as

Table 7
Default model—failing to predict failure.
We report estimates from a baseline default model estimated for low and full-documentation loans originated from 1997 to 2000 in Panel A. A loan is defined to be in default if it is delinquent for at least 90 days within 24 months from the year of origination. Panel B reports the β coefficients from a regression of prediction error on FICO score and LTV ratio for loans issued from each year 2001 to 2006, and also reports the mean prediction errors for each vintage. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

Panel A: Coefficients of baseline model in low securitization regime, 1997–2000

FICO	*r*	*LTV*	$I^{Low} \times$ *FICO*	$I^{Low} \times$ *r*	$I^{Low} \times$ *LTV*	Pseudo R^2 (percent)	Number of observations
−0.009***	0.231***	0.003***	0.001***	−0.043***	−0.008***	7.05	267,511
(0.0001)	(0.006)	(0.001)	(0.0001)	(0.016)	(0.001)		

Panel B: Prediction errors during high securitization regime.

Origination year	Actual and predicted defaults		Regression of prediction error on FICO, LTV			
	Mean prediction error (percent)	Actual defaults (percent)	β_{FICO} ($\times 10^{-3}$)	β_{LTV} ($\times 10^{-2}$)	Number of observations	Adjusted R^2 (percent)
2001	3.96***	16.0	−0.123***	0.052***	128,772	0.05
				(0.018)	(0.010)	
2002	4.70***	14.1	−0.197***	0.082***	152,057	0.15
				(0.015)	(0.010)	
2003	5.01***	11.9	−0.428***	0.077***	308,340	0.61
				(0.010)	(0.010)	
2004	7.79***	13.9	−0.621***	0.061***	596,485	0.97
				(0.008)	(0.004)	
2005	14.67***	21.1	−1.341***	0.143***	788,299	3.90
				(0.030)	(0.007)	
2006	25.49***	33.2	−1.120***	0.190***	608,559	1.60
				(0.012)	(0.005)	

$$\textit{Predicted Default}_{i,t+k} \equiv \text{Prob}(\widehat{\text{Default}}_{i,t+k}=1), \text{ where}$$

$$\text{Prob}(\widehat{\text{Default}}_{i,t+k}=1) = \Phi(\hat{\beta}_{1,t}\cdot X_{i,t+k} + \hat{\beta}^{Low}_{1,t}\cdot I^{Low}_{i,t+k}X_{i,t+k}). \quad (10)$$

We then examine the actual default experience of loans issued in each of years 2001 to 2006. The prediction error is computed as $\textit{Prediction Error}_{i,t+k} = \textit{ActualDefault}_{i,t+k} - \textit{Predicted Default}_{i,t+k}$.

In the second and third column under Panel B of Table 7, we report the mean prediction error and the actual proportion of loans in default in each year. As can be noted from the table, the mean prediction error is positive (and significantly different from zero at the 1% level) throughout. For loans issued in the period 2001–2004, the mean prediction error amounts to 25–50% of the actual default proportion, and climbs even higher for 2005 and 2006 loans. The increasing size of the prediction error indicates that the fit of the model worsens over time.

If there is systematic underprediction at low FICO scores and high LTV ratios, the prediction error should decline in magnitude as the FICO score increases and LTV ratio falls. To check this, we estimate yearly the ordinary least squares regression for loan i in year $t+k$ (where $t=2000$ and $k=1,2,\ldots,6$) as

$$\textit{Prediction Error}_{i,t+k} = \alpha + \beta_{FICO}\times FICO_{i,t+k} + \beta_{LTV}\times LTV_{i,t+k}.$$

The last four columns of Panel B of Table 7 report the coefficients on the FICO scores and LTV ratio for loans issued in each of the years 2001 to 2006. As can be observed from Columns 2 and 3, the coefficient β_{FICO} is negative and β_{LTV} is positive and significant across 2001 to 2006. The magnitudes seem large. For instance, an increase in 1 standard deviation in the FICO score (about 70 points) leads to a reduction in the prediction error of about 33.5% for 2006 loans. Similarly, a 1 standard deviation increase in LTV ratio (about 10%) leads to a reduction in prediction error of about 9.4% for 2006 loans.

We have shown that the mean prediction error is positive. As further confirmation, we plot the Epanechnikov kernel density of mean prediction errors over time.[16] If the relation between defaults and observables has not changed since the baseline period, one would expect the average of the mean prediction error across the entire sample to be approximately zero. Positive macroeconomic shocks should shift the distribution to the left; that is, there would be fewer defaults than expected, so prediction errors would be negative. Negative macroeconomic shocks should shift it to the right, with positive prediction errors. As is clear from Fig. 1, the distributions show that on average the mean prediction error has been positive in each year. Moreover, the distribution of the mean prediction error progressively shifts to the right over time, as securitization becomes more prevalent in the subprime market. It is striking that the vast majority of prediction errors are positive in each year and remarkably few observations have negative mean prediction errors. Importantly, we observe this phenomenon even in years in which the economy was doing well and house prices were increasing (specifically, for loans issued between 2001 and 2004).

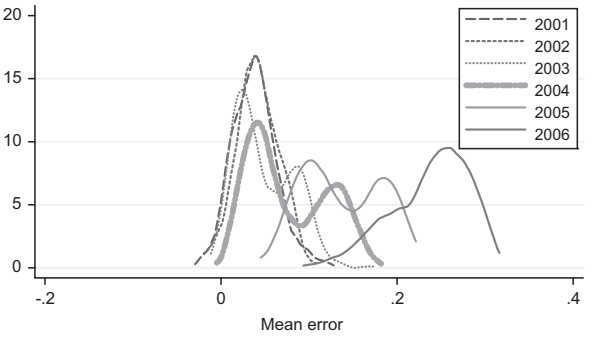

Fig. 1. Kernel density of mean prediction errors over time, all loans. Kernel density of mean prediction errors (*Actual Defaults − Predicted Defaults*) of a baseline model estimated for loans issued in 1997 to 2000. For each subsequent year, we first determine the mean prediction error at each FICO score and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991).

Our test above estimates the coefficients of the model in the window 1997 to 2000 and considers the prediction errors in the period 2001 to 2006. As seen from Table 2, a steady increase is evident in securitization over the latter period. Hence, an alternative way to conduct this test is to use as much historical data as available for each year to tease out the incremental effect of additional securitization on the prediction errors of a default model. Using a rolling window, we predict defaults for loans issued in years 2005 and 2006, which allows the baseline model to include a few years of data from the high securitization regime. Thus, we expect the prediction errors to be smaller. For 2005 loans, the baseline model is estimated over the period 1997 to 2004, and for 2006 loans the base period is 1997 to 2005.[17] The results are qualitatively similar, though the magnitudes of the errors are reduced. The average prediction error in this specification is 8.3% for 2005 loans (compared to 14.7% in the baseline specification) and 15.1% for 2006 loans (compared to 25.5% in the baseline specification).

Our results are also robust to the introduction of lender fixed effects in the baseline regression model in Eq. (9). We re-estimate the model adding lender fixed effects for the largest seven hundred or so lenders, which comprise 90% of securitized loans over the entire sample period. The results on prediction errors are essentially similar to those reported in Table 7 and shown in Fig. 1. For brevity, these results are not reported in the paper. The important conclusion is that our results on defaults are also not driven

[16] These plots are constructed as follows. For each year, across all loans at each FICO score, we determine the mean prediction error. We then plot the kernel density using the mean errors at each FICO score. The plots look similar if the errors are weighted by the actual number of loans at each FICO score.

[17] This is a stringent specification. We track default on loans issued in 2004 until the end of 2006 and on loans issued in 2005 until the end of 2007. As a result, the rolling window estimation incorporates adverse forward information in the baseline model. Consequently, the errors we obtain from such a model are smaller than those a regulator could obtain using data only available in real time.

Table 8
Default model—placebo test.
We report estimates from a baseline default model estimated for low-documentation loans issued in 1997 and 1998 in Panel A. A loan is defined to be in default if it is delinquent for at least 90 days within 24 months from the year of origination. Panel B reports the β coefficients from a regression of prediction error on FICO score and LTV ratio for loans issued in 1999 and 2000, and also reports the mean prediction errors for each vintage. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

Panel A: Coefficients of baseline model in low securitization regime

Origination years	*FICO*	*r*	*LTV*	Pseudo-R^2 (percent)	Number of observations
1997–1998	−0.009***	0.249***	−0.008***	8.11	16,002
	(0.0005)	(0.034)	(0.003)		
1997–1999	−0.007***	0.259***	−0.003*	7.94	33,868
	(0.003)	(0.022)	(0.001)		

Panel B: Prediction errors during high securitization regime.

Origination	Actual and predicted defaults		Regression of prediction error on FICO, LTV			
year	Mean prediction error (percent)	Actual defaults (percent)	β_{FICO} ($\times 10^{-3}$)	β_{LTV} ($\times 10^{-2}$)	Number of observations	Adjusted R^2 (percent)
1999	0.91	11.0	0.039	0.026	17,866	0.01
			(0.038)	(0.023)		
2000	0.97	11.9	0.039	−0.026	24,591	0.01
			(0.034)	(0.020)		

by a change in lender composition over the sample period, but instead hold within each lender.[18]

5.3. Placebo test: predictability of defaults in low securitization regime

Across different years in the low securitization regime, there should be no substantive change in a lender's incentives to collect information about a borrower or property. Thus, the mapping between observables and defaults should be approximately similar from year to year. This argument forms the basis of a placebo test in which we assess whether a default model estimated during a low securitization regime generates small prediction errors in another period with relatively low securitization.

To conduct the test, we predict defaults on low-documentation loans issued in 1999 and 2000, using a baseline model estimated from 1997 and 1998 for 1999 loans and 1997 through 1999 for 2000 loans (i.e., employing a rolling window). The results are reported in Table 8. As shown in Panel B, the mean prediction error is not significantly different from zero. Further, when we regress the prediction errors on FICO score and LTV ratio for each year 1999 and 2000, the β_{FICO} and β_{LTV} coefficients are insignificant, in contrast to the results in Table 7.

In Fig. 2, we plot the kernel distribution of the mean prediction error at each FICO score. In contrast to Fig. 1, the mean errors are centered around zero, suggesting that there is no systematic underprediction by the baseline model. Thus, the control test is consistent with our hypothesis.

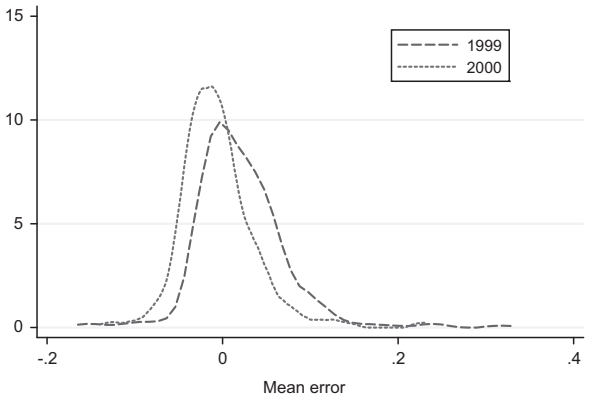

Fig. 2. Placebo test—mean prediction errors in low securitization regime. This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults* − *Predicted Defaults*) for low-documentation loans issued in 1999 to 2000. The baseline model for 1999 loans is estimated over 1997 and 1998 and the baseline model for 2000 loans is estimated from 1997 through 1999. For each year 1999 and 2000, we first determine the mean prediction error at each FICO score and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991).

6. Cross-sectional tests

We now describe several cross-sectional tests that both confirm our findings and alleviate the concern that some of our results on prediction errors could be due to macro factors other than securitization levels that also changed over time.

6.1. Full- and low-documentation loans

To directly test that our results are driven by the information channel, we separately consider low- and full-documentation loans. More information remains unreported on low-documentation loans, compared with

[18] We obtain similar results when we consider only fixed rate mortgages, which confirms that our results are not driven by a change in loan composition over time. In addition, we perform the same exercise on loans issued by NCFC, and obtain qualitatively similar results.

full-documentation loans. Thus, all else equal, a default model fitted during a low securitization era should perform relatively better (in terms of default predictions in the high securitization period) on full-documentation loans. Importantly, the distribution of full- and low-documentation loans across zip codes is similar. To check this, we sort the volume of each kind of loan by zip code over 2001–2006, and consider the top 25% of zip codes in each case (which contribute over 60% of the volume of each kind of loan). A large proportion of zip codes (about 82%) are common across the two lists. Thus, under the assumption that low- and full-documentation borrowers are equally sensitive to changes in the economy, any differential effects across the two kinds of loans are insulated from macroeconomic and shocks at the zip code level to employment and house prices.

To evaluate how prediction errors vary across the two kinds of loans, we use a rolling window specification and fit separate baseline models for full- and low-documentation loans. That is, for predicting default probabilities on loans issued in year $t+1$, the baseline model is estimated over years 1 through t, where year 1 is 1997. For each kind of loan $s = Low, Full$, the baseline specification is a logit model of the form

$$\text{Prob}(\text{Default}_i^s = 1) = \Phi(\beta_{1,t}^s \cdot X_i^s),$$

where the vector X_i is the same as described earlier in this section. Let $\hat{\beta}_{1,t}^s$ be the estimated coefficients from this regression. The predicted default probability for loans issued in year $t+1$ is then estimated as

$$\text{Prob}(\widehat{\text{Default}}_{i,t+1}^s = 1) = \Phi(\hat{\beta}_{1,t}^s \cdot X_{i,t+1}^s),$$

Panels A and B of Fig. 3 plot the Epanechnikov kernel density of mean prediction errors at each FICO score over time separately for full- and low-documentation loans. The plots suggest that, as predicted, the prediction errors are larger for low-documentation loans than for full-documentation loans. The mean prediction errors for full- and low-documentation loans are reported in Table 9, and are substantially greater than those reported for the years 1999 and 2000 in Panel B of Table 8. For loans issued in 2003 and later, the mean errors are approximately 80% higher for low-documentation loans.

6.2. Loans across bordering zip codes of states with different foreclosure regulations

In our next cross-sectional test, we exploit differences in the ease of securitization induced by different foreclosure regulations across states. As highlighted by Pence (2006), some states require judicial foreclosure; that is, a foreclosure must take place through the court system. In contrast, other states have a nonjudicial procedure in which a lender has the right to sell a house after providing only a notice of sale to the borrower. A judicial foreclosure imposes substantial costs, including time delay, on a lender.

Table 9
Mean prediction errors for low- and full-documentation loans.

We report the mean prediction errors for low- and full-documentation loans issued from 2001 through 2006. The estimation uses a rolling window approach with separate baseline models for low-documentation and full-documentation loans. That is, the predictions for year $t+1$ are based on a model estimated over the years 1 through t, where year 1 is 1997. *** indicates significance at the 1% level; ** at the 5% level; * at the 10% level.

Origination year	Low-documentation (percent)	Full-documentation (percent)	Difference (percent)
2001	3.40	3.80	−0.40
2002	2.78	2.79	−0.01
2003	3.20	2.21	0.99***
2004	5.17	3.51	1.66***
2005	10.58	5.85	4.73***
2006	20.11	9.84	10.27***

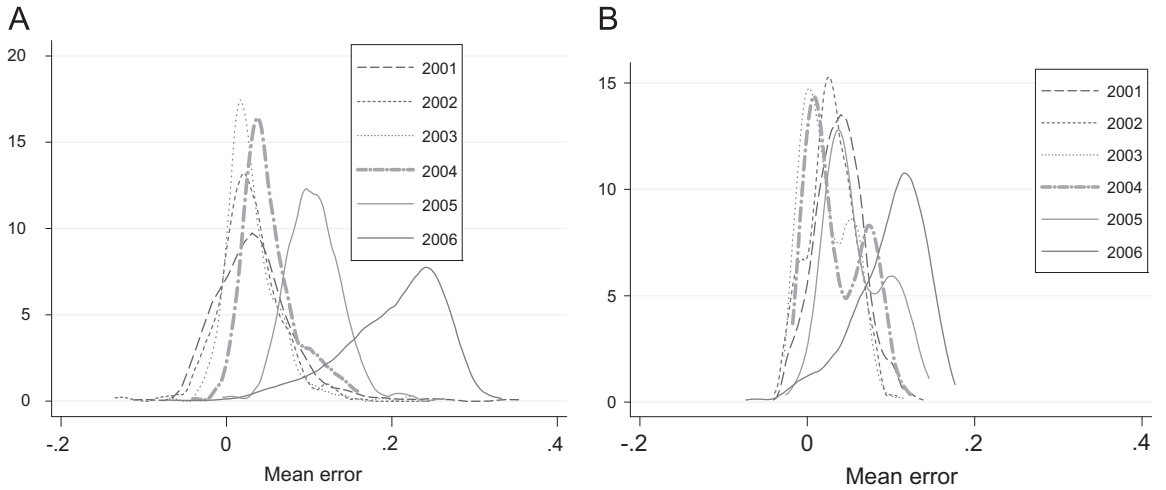

Fig. 3. Mean prediction errors for low- and full-documentation loans. This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults* − *Predicted Defaults*) on low-documentation (Panel A) and full-documentation (Panel B) loans of a baseline model using a rolling estimation window. The prediction errors for year $t+1$ are from a baseline model estimated over 1997 to year t. For each year, we first determine the mean prediction error at each FICO score and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991).

We postulate that the ease of securitization is higher in states with nonjudicial foreclosure. Following the arguments in Pence (2006), the supply of securitized mortgage credit could be lower in states with judicial foreclosures. Moreover, the distance between borrower and investor created by securitization represents a more significant wedge when the foreclosure proceedings are more complicated. As a result, it is relatively more costly for dispersed investors to renegotiate with a delinquent borrower (see Piskorski, Seru, and Vig, 2010; Agarwal, Amromin, Ben-David, Chomsisengphet, and Evanoff, 2011) or to initiate judicial proceedings. We confirm empirically that securitization indeed appears to be easier in states with nonjudicial foreclosure. Our prediction on the default mapping then implies that a historical default model would breakdown more for loans in nonjudicial states. That is, the prediction errors from a default model fitted to past data should be higher for loans in nonjudicial foreclosure states.

To account for the fact that economic conditions across the two sets of states can vary more broadly, we adopt the border strategy used by Pence (2006) and Mian, Sufi, and Trebbi (2011). That is, we identify and match counties on either side of a state border that are otherwise comparable to each other. In particular, we begin with Metropolitan Statistical Areas (MSAs) that cross state lines. For counties in these MSAs, we determine the population (from the 2000 census), median income, and the percent of the population below the poverty line, younger than 40, with a high school diploma, and with a higher education degree. For two counties in different states to be considered a match, the demographic variables listed above must be within 1 standard deviation of each other. We find a unique pair of counties in each MSA across state lines that satisfy the above criteria. Finally, we consider only loans made in the zip codes of this matched sample of counties (see Pence, 2006 or Mian, Sufi, and Trebbi, 2011 for more details).

Panels A and B of Fig. 4 show the number of new securitized loans per thousand households in the control and treatment group over time in our main sample. We exhibit the data for low-documentation loans in Panel A and full-documentation loans in Panel B. After 2002, a clear divergence emerges in the number of securitized loans of both types in states with nonjudicial foreclosures. The gap between states with nonjudicial and judicial foreclosures increases until 2006, coinciding with the overall securitization boom in subprime mortgage loans. Combining the two sets of loans yields a similar trend. We therefore consider loans in states with nonjudicial foreclosures as the high securitization (or treatment) group, and loans in states with judicial foreclosures as the low securitization (or control) group. Panels C and D of Fig. 4 plot the kernel densities of interest rates for low-documentation and full-documentation loans respectively, in both judicial and nonjudicial foreclosure states. We compute the average interest rate at each FICO score to plot these densities. As the figure shows, the interest rate distributions are very similar across the two kinds of states.[19]

We repeat the analyses of Sections 4.1 and 5.2 on the matched sample of loans. First, we separately consider loans in judicial and nonjudicial foreclosure states and regress the interest rate on the borrower's FICO score and the LTV ratio for the period 1997–2000 and for each issue year from 2001 to 2006. For both sets of states, the R^2 of the regression increases over time as securitization increases, displaying a similar pattern to that in Table 3. The R^2 starts out lower for nonjudicial foreclosure states relative to judicial foreclosure states, but becomes larger during the boom period (2004–2006). Specifically, the R^2 increases from 0.107 in 1997–2000 to 0.341 in 2005 and 0.429 in 2006 for loans in judicial foreclosure states and from 0.073 in 1997–2000 to 0.399 in 2005 and 0.468 in 2006 for loans in nonjudicial foreclosure states. That is, the R^2 of the regression increases by a greater amount in the nonjudicial foreclosure states, consistent with a greater reliance on hard information in those states. The coefficients on the borrower FICO score and LTV ratio are similar to those in Table 3 in both cases. For brevity, we do not report the details.

Next, we estimate the statistical default model specified in Eq. (8) above for the period 1997–2000. We estimate the model separately for low- and full-documentation loans in each of the two kinds of states. We then determine the default prediction errors for loans issued in each year from 2001 to 2006. For brevity, in Fig. 5 we show the kernel densities of the average across years of the mean prediction error at each FICO score. We separate out states with judicial and nonjudicial foreclosure proceedings and low- and full-documentation loans. The figure shows that, as expected, the prediction errors are positive in both sets of states. However, for both low- and full-documentation loans, the density of the errors is shifted to the right for loans in nonjudicial foreclosure states (the states with larger levels of securitization). The shift is more pronounced for low-documentation loans, for which soft information is likely to be more important. The overall mean prediction error is 0.1 in states with judicial foreclosures and 0.122 in states with nonjudicial foreclosures. The error is, therefore, over 20% greater in the latter states, and the difference is statistically significant.

This cross-sectional test supports our main finding that statistical default models perform especially poorly when the levels of securitization are high. Our test here is stringent: We compare both low- and full-documentation loans in a matched set of counties that lie on different sides of a state border but are otherwise comparable on several observables. Even in this narrow range of counties, the connection between securitization and defaults remains, and our results are stronger for low-documentation loans.[20]

[19] An alternative way to condition on the FICO score and LTV ratio is to plot the residuals from the regression of interest rates on FICO score and LTV ratio. The resulting kernel density plots are again similar.

[20] One concern could be that in states with nonjudicial foreclosures an alternative economic force drives both a lack of screening by lenders and ease of securitization. For instance, suppose the fact that collateral can be seized more easily by a lender if a loan defaults in a nonjudicial foreclosure state makes lenders less likely to screen loans in this state. Reduced information collection by lenders could also increase the ease of securitization, because investors are less likely to face adverse selection. This would induce a correlation between lower screening and ease of securitization. However, this argument cannot explain the effects we find in the time series. Over time, although the foreclosure laws across states have been stable, we find an increase in the ease of securitization and a

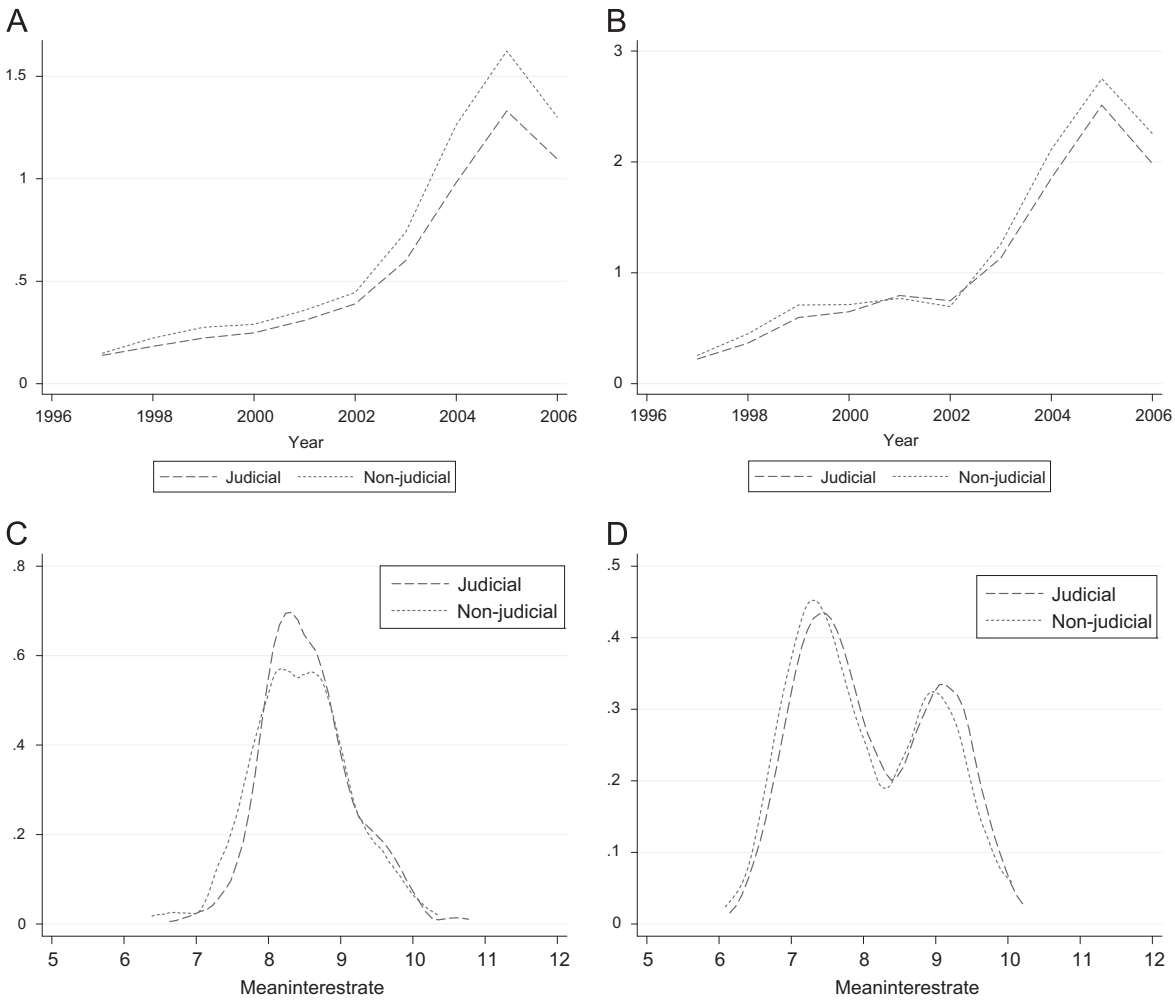

Fig. 4. Loans in judicial and nonjudicial foreclosure states. The average annual number of newly-issued securitized loans per thousand households are shown for low-documentation loans (Panel A) and full-documentation loans (Panel B) in states that require judicial foreclosures and those that allow nonjudicial foreclosures. The Epanechnikov kernel densities of interest rates are shown for low-documentation loans (Panel C) and full-documentation loans (Panel D) in states that require judicial foreclosures and those that allow nonjudicial foreclosures. To generate the kernel densities, we first average the interest rate at each FICO score. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991).

6.3. *Low-documentation loans on either side of a FICO score of 620*

The previous two tests consider borrowers across the FICO spectrum. For our third test, we consider a cross-section of borrowers in a narrow range of FICO scores who are similar in terms of their observable characteristics but exogenously differ in the likelihood that their loans will be securitized. Following guidelines set by the Federal National Mortgage Association (FNMA) and the Federal Home Loan Mortgage Corporation (FHLMC) in the mid-1990s, a FICO score of 620 has become a threshold below which it is difficult to securitize low documentation loans in the subprime market. Keys, Mukherjee, Seru, and Vig (2010) and Keys, Seru, and Vig (2012) show that the ease and likelihood of securitization is greater for low-documentation loans with FICO scores just above 620 (call these 620^+ loans) compared to those with FICO scores just below 620 (620^- loans). Importantly, other observable borrower and loan characteristics are the same across the two sets of loans. This allows us to construct a cross-sectional test for borrowers within the low-documentation market.

Our test compares the prediction errors on 620^+ low-documentation loans with those on 620^- low-documentation loans, where 620^+ includes FICO scores from 621 to 630 and 620^- includes FICO scores from 610 to 619. For brevity, we conduct this test averaging the prediction errors (at each FICO score) for all low-documentation loans issued in the period 2001–06. The baseline model used is the model in Eq. (9), estimated on only 620^+ and 620^- loans. The results are shown in Fig. 6. The prediction errors are indeed lower for 620^- loans (16.6%) than 620^+ loans (18.2%). The difference in mean errors of 1.6% is statistically significant at the 1% level.

(*footnote continued*)
corresponding increase in the extent of underprediction by a statistical default model.

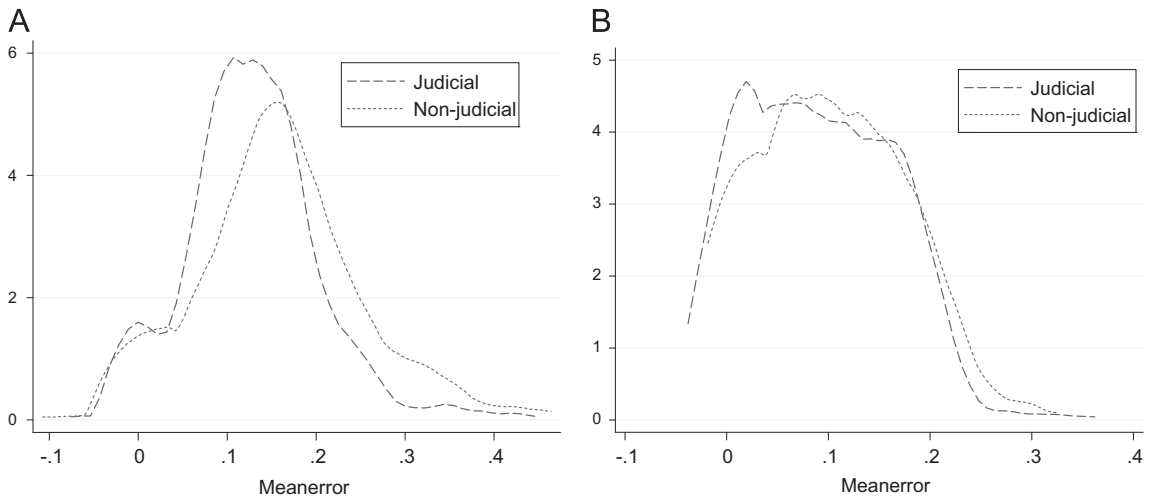

Fig. 5. Mean prediction errors for loans in judicial and nonjudicial foreclosure states. This figure shows the Epanechnikov kernel density of mean prediction errors (*Actual Defaults − Predicted Defaults*) of a baseline model estimated for loans issued in 1997 to 2000. The model is estimated separately for low- and full-documentation loans in states with judicial and nonjudicial foreclosures. For each subsequent year, we first determine the mean prediction error at each FICO score. We then average the errors across years at each FICO score and plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991). (Panel A) Low-documentation loans and (Panel B) Full-documentation loans.

7. Alternative hypotheses

An important alternative hypothesis is that our finding of a positive prediction error in default models is driven by falling house prices. We provide three pieces of evidence to rule this out. First, in each of our cross-sectional tests, the two sets of loans being compared are subject to the same effects of changing house prices. Therefore, changing house prices cannot explain the differences across the loans being compared in each case. Second, in Fig. 1, we show positive prediction errors from a statistical default model even for loans issued in the period 2001–2004. For these loans, house prices were increasing in the relevant period of two years beyond issue. Only in August 2007 did the composite (i.e., national level) Case-Shiller index indicate a fall from its value 24 months earlier.[21] Undoubtedly, a fall in house prices is partly responsible for the surge in defaults for loans issued in 2005 and 2006 (see, for example, Mayer, Pence, and Sherlund, 2009; Mian and Sufi, 2009). Third, in Section 7.1, we show that our results are qualitatively robust to the explicit inclusion in the default model of future changes in house prices.

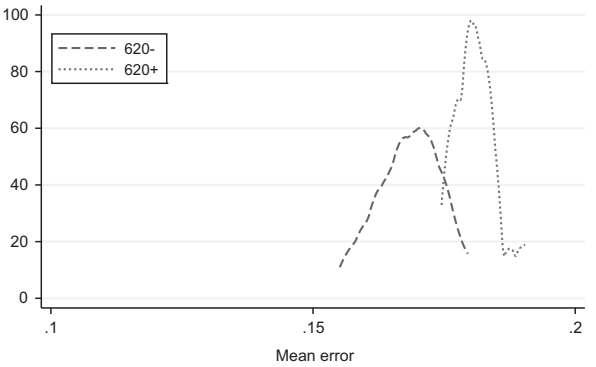

Fig. 6. Mean prediction errors for 620+ and 620− low-documentation loans. This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults − Predicted Defaults*) for low-documentation loans issued in 1999 to 2000. The dashed line represents the prediction errors on loans with FICO scores from 621 to 630 (i.e., 620^+ loans) and the dotted line the prediction errors on loans with FICO scores from 610 to 619 (i.e., 620^- loans). We determine the prediction errors across loans at each FICO score and then plot the kernel density of the mean errors. The prediction errors are relative to a baseline model estimated on loans issued in the years 1997 through 2000. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991).

7.1. Explicitly accounting for the effect of changing house prices

For each loan, we construct a house price appreciation (*HPA*) variable as follows. We begin with the state-level quarterly house price index constructed by the Office of Federal Housing Enterprise Oversight. For each state s, a house price index for each year t, $h_{s,t}$, is constructed as a simple average of the indices over four quarters. Consider loan i issued in state s in year t. The house price appreciation variable for loan i is set to the growth rate of house prices over the next two years, $HPA_i = (h_{s,t+2} - h_{s,t})/h_{s,t}$. We include HPA_i in the vector of loan characteristics X_i in both the baseline and predictive regressions. Our specification is stringent: It clearly includes more information than available to an econometrician at the time the forecast is made and soaks up more variation in defaults than a prediction made

[21] Two possible explanations exist for borrowers defaulting when house prices increase. First, over 70% of the loans in our sample have a prepayment penalty, increasing the transaction cost to a borrower of selling the house. Second, some borrowers who experience an increase in home prices could be taking out additional home equity loans, effectively maintaining a higher LTV ratio than reported in the sample. The latter effect is consistent with our story, because information on whether a borrower could be credit-constrained in the future and take out additional home loans is soft information potentially observable by a lender but not reported to the investor.

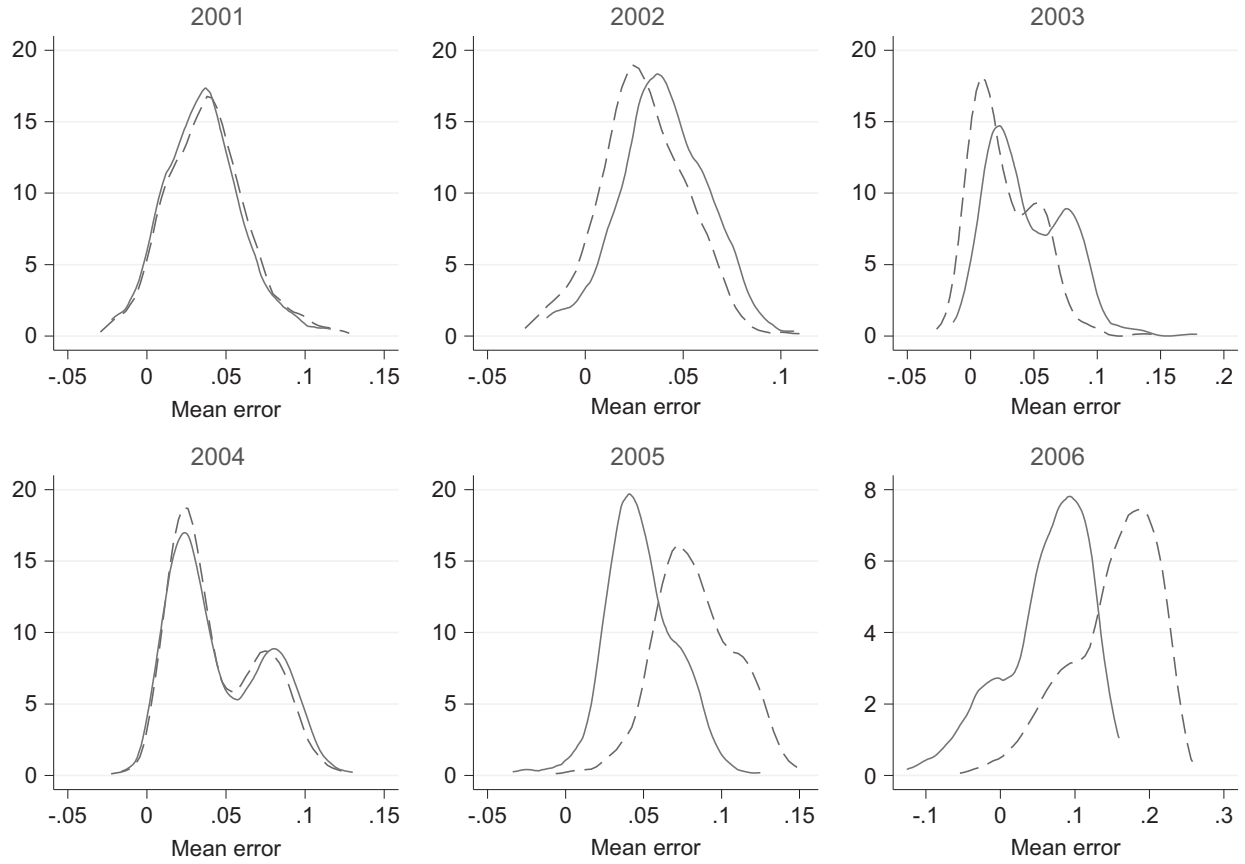

Fig. 7. Explicitly incorporating house price effects. This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults* − *Predicted Defaults*) on all loans of a baseline model using a rolling estimation window. The prediction errors for year $t+1$ are from a baseline model estimated over 1997 to year t, with and without including house price appreciation (*HPA*) as an explanatory variable. For each year, we first determine the mean prediction error at each FICO score and then plot the kernel density of the mean errors. For each year, the dashed line represents the density of errors without *HPA*; the solid line, the density of errors with *HPA* included. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991).

in real time (in other words, the specification assumes the regulator or rating agency has perfect foresight).

We re-estimate the baseline model of Eq. (9) after including the *HPA* variable (both by itself and interacted with I^{Low}, the low-documentation dummy) on the right-hand side. We then predict default probabilities for loans issued in each of the years 2001 through 2006 using Eq. (10) after including the *HPA* variable and its interaction with I^{Low}. A rolling window is used for this estimation, so default probabilities for loans issued in year $t+1$ are predicted based on coefficients estimated over years 1 through t, where year 1 is 1997. In Fig. 7, we plot the Epanechnikov kernel density of mean prediction errors (computed at each FICO score) in each year 2001 through 2006. For ease of comparison, the figure has six panels, each panel showing the kernel density of mean out-of-sample prediction errors in a given year with and without including house price appreciation as an explanatory variable, using a rolling estimation window in each case.

Two observations emerge from the figure. First, for 2001–2004 loans, there is not much difference in the two kernel densities. In fact, for 2002–2003 loans, including the house price effect slightly magnifies the prediction errors. Second, the prediction errors for loans issued in 2005 and 2006 are reduced in magnitude when the effect of house prices is included. In particular, using a rolling window for estimating the baseline model, the mean prediction error for 2005 loans falls from 8.3% to 4.9% when *HPA* is included as an explanatory variable, and for 2006 loans it falls from 15.1% to 6.1%. Thus, for these two years, approximately 50% of the mean prediction error survives over and above the effect of falling house prices. Therefore, even after accounting fully for the effect of falling house prices on defaults, the prediction errors exhibit patterns consistent with our predictions. It continues to be striking how few of the mean errors are less than zero across the entire period 2001–2006.[22]

These results are consistent with those of Gerardi, Lehnert, Sherlund, and Willen (2008), who suggest that a

[22] In unreported tests, we repeat the analysis for low- and full-documentation loans after including the house price effect. For loans issued in 2001–2004, the results are similar to those reported in the cross-sectional test described earlier. For loans in 2005 and 2006, the magnitudes of the prediction errors are reduced for both groups of loans, but the errors continue to be larger for low-documentation loans.

statistical model of foreclosures fitted over the period 2000–2004 performs well on predictions for loans issued in the period 2005–2007 once perfect foresight of house price trajectories is taken into account. It is important to note that their default event is a foreclosure (i.e., a lender taking legal steps to seize a house from a borrower) whereas ours is a 90-day delinquency (i.e., a borrower being 90 days behind on loan payments). Many delinquencies lead to renegotiation instead of foreclosure (see Piskorski, Seru, and Vig, 2010), so the number of foreclosures is substantially less than the number of delinquencies. Once we account for this factor, our results are broadly similar to theirs in terms of the quantitative magnitudes. For example, for 2005 loans, they find prediction errors of approximately 2% on foreclosures in a 24-month window from issue, whereas we find an approximately 4% prediction error on delinquencies.

For robustness, we repeat our analysis on the effect of house prices on default prediction errors using a few different specifications. First, in both the baseline model of Eq. (9) and the predictive model in Eq. (10) we interact the *HPA* variable with the borrower FICO score, the LTV ratio, and the interest rate, to allow for the possibility that the effect of house price appreciation could vary depending on other characteristics of the borrower or loan. Second, we use data at the zip code level from CoreLogic to determine the *HPA* variable for each loan and re-estimate the prediction errors. In both cases, the kernel densities of prediction errors are very similar to the solid lines shown in Fig. 7. For brevity, we do not show these figures in the paper.

In sum, incorporating house price changes substantially reduces the prediction errors from the default model for loans issued in 2005 and 2006. However, the mean prediction error remains large, and the kernel densities show that the prediction errors at each FICO score are consistently positive. Further, we continue to find positive prediction errors in the period 2001–2004, when house prices were increasing.

7.2. Other alternatives

We briefly consider a few other alternative explanations for our results. Many of these explanations are effectively ruled out by our cross-sectional tests. Particularly on the test comparing loans across state borders and the test comparing loans on either side of a FICO score of 620, the two groups of loans are similar on several dimensions. Any factor other than ease of securitization can be a plausible alternative explanation only if it differentially affects the two groups of loans being compared in each test.

7.2.1. Lower cost of capital

One benefit of securitization is a lower cost of capital for the lender. As the cost of capital falls, some risky borrowers who represent negative NPV projects at the higher cost of capital now become positive NPV projects. Thus, given a set borrowers with the same FICO score, the lender naturally makes loans to more risky borrowers at the lower cost of capital. Therefore, as securitization increases, the quality of loans issued worsens, leading to positive prediction errors from a statistical default model.

However, the lower cost of capital channel has a very different prediction on the interest rate distribution, compared to our channel of loss of soft information. Even at a lower cost of capital, there should be a difference in the interest rates charged to a more risky and a less risky borrower. Thus, over time, if the pool of issued loans includes borrowers with greater risk, the dispersion of interest rates at a given FICO score should increase. However, as shown in Section 4.1, the dispersion of interest rates falls as securitization increases, especially at low FICO scores. This pattern is consistent with a loss of soft information for low hard information signals, but not the riskier borrowers channel.[23]

7.2.2. Second-lien mortgages taken out by borrowers

Suppose that, after a loan has been issued, a borrower then obtains a second-lien loan on the property. The borrower reveals himself to be financially constrained and is more likely to default in the future. It is well known that there was an increase in second-lien loans in the subprime mortgage sector over the period of our study. Could our results on defaults be driven by this factor?

In our data, information on second-lien loans or on the cumulative loan-to-value (CLTV) ratio across all loans is not reliably available in the earlier part of the sample. Regardless, with the information that is available, we re-estimate our default model using this variable and find that the results on underprediction are similar to those reported in the paper.

7.2.3. Standardization of mortgage terms

In Table 4, we demonstrate a shrinkage in the distribution of interest rates over time. During our sample period, there was some standardization of terms of mortgage loans for transparency reasons (see Kroszner, 2007), which could explain such shrinkage. However, we also find that the shrinkage occurs in significantly greater amounts for borrowers at low FICO scores. This cannot be explained by the standardization of contractual terms unless the loan terms were already standardized at high FICO scores by 1997. Further, standardization per se cannot explain the patterns we find on the changed mapping between observables and default rates.

8. Role of investors

The boom in securitization of subprime mortgage loans over our sample period was possible only because investors showed a continued and increasing willingness to buy these loans. Our analysis shows that a naïve regulator relying only

[23] Additional evidence against the cost of capital channel is provided by Keys, Mukherjee, Seru, and Vig (2010), who conduct a cross-sectional test using similar data, and show that defaults on a portfolio that is more likely to be securitized exceed defaults on a portfolio that has similar risk characteristics but is less likely to be securitized. Their test to rule out the cost of capital channel also involves the dispersion of the interest rate distribution.

on past data would underestimate loan defaults. Were investors better able to forecast defaults?

Our analysis is largely agnostic on whether investors were also fooled by the change in the lending regime. Importantly, our predictions obtain even when both lenders and investors are fully rational, as formally demonstrated in a theoretical model by Rajan, Seru, and Vig (2010). Because soft information cannot be contracted upon, a moral hazard problem is created: Investors cannot provide lenders with an incentive to collect it. As a result, in equilibrium, in a high-securitization regime lenders do not collect soft information, so the quality of the loan pool worsens relative to a low-securitization regime. Lenders are aware of this, anticipate higher defaults on loans, and price them accordingly.

This effect is exacerbated if investors are boundedly rational and price loans using default predictions from a naïve method. Loan prices will then be too high, especially for borrowers on whom the unreported information is an important predictor of quality. Lenders now have an even stronger incentive to ignore the unreported information in approving loans and setting interest rates. As a result, the tendency of a statistical model to underpredict defaults for these borrowers worsens.

Empirically, it is important to consider whether investors rationally anticipated the increase in defaults implied by our results: With rational investors, asset prices can be used to fine tune regulation. A direct test of investor rationality is difficult to conduct. We do not have data on the pricing of CDO tranches backed by subprime mortgage loans. As an indirect test, we consider the subordination levels of AAA tranches for new non-agency pools consisting of loans originated in 2005 and 2006. We have already shown (Figs. 1 and 7) that a statistical default model most severely underestimates actual defaults in 2005 and 2006. The subordination level measures the magnitude of losses an equity tranche can absorb, before the principal of the AAA tranches is at risk. Thus, if rating agencies were correctly forecasting future defaults, the subordination levels in the pools must have a positive correlation with the prediction errors of the default model (otherwise the tranches should not have been rated AAA).

To highlight whether a relationship exists between subordination levels of AAA tranches and prediction errors on default, we consider only pools for which prediction errors (i.e., actual defaults minus predicted defaults given the baseline model) are likely to be high. In particular, we restrict attention to pools with at least 30% low-documentation loans. Subordination level information is obtained from Bloomberg and cross-checked with information provided in the Intex database. We compute prediction errors using the coefficients from the baseline default model in Eq. (9). We omit the details for brevity. At best, we find a weak relation between the subordination level of AAA-tranches and the mean prediction error on the pool, suggesting that rating agencies were unaware of or chose to overlook the underlying regime change in the quality of loans issued as securitization increased.

These results are consistent with the work of Ashcraft, Goldsmith-Pinkham, and Vickery (2010), who find that during this period subordination levels do not adjust enough to reflect the increased riskiness of originated loans. Similarly, Benmelech and Dlugosz (2009) and Griffin and Tang (2012) argue that ratings of CDO tranches were aggressive relative to realistic forward-looking scenarios. More directly, Coval, Jurek, and Stafford (2009) consider the pricing of CDO tranches backed by credit default swaps, and conclude that the spreads are much lower than those available in other asset markets for similar risks. Along similar lines, Faltin-Traeger, Johnson, and Mayer (2010) find that the ability of spreads to predict future downgrades on asset-backed security tranches is weak. Therefore, evidence suggests that some classes of structured products and subprime-backed securities were mispriced by investors.[24]

Investors could have been overly optimistic about the path of future house prices, as indicated by Gerardi, Lehnert, Sherlund, and Willen (2008). Investors care about losses rather than defaults per se, and with rising house prices small prediction errors in default models could result in only small losses to investors. Like we do in Section 7.1, Gerardi, Lehnert, Sherlund, and Willen (2008) consider perfect foresight of house price trajectories. Even after accounting for this trajectory of house prices, their model underpredicts foreclosures. The errors are, in fact, in the same range as obtained in our exercise. Knowing the actual expectations of investors with respect to future house prices is not possible. Nevertheless, perfect foresight is a conservative benchmark. Any realistic model of house price changes (based on expectations at that time) would yield a smaller house price decline for loans issued in the years 2005 and 2006 compared with perfect foresight, and so to a more severe underestimation of defaults.

8.1. *Securitization as a repeated game*

The process of securitizing mortgage loans is a repeated game, with an issuer continually generating loans that are then sold to investors. In such a context, one can imagine that there are dynamic disciplining mechanisms: If defaults in any one year are too high, investors can punish an issuer by not buying its loans in the future.

Our analysis is silent on why investors did not react negatively to the rising defaults for loans issued in the years 2001–2004. In practice, investors care about the overall loss suffered on a package of loans rather than the default rate per se. In an era with rising house prices, even if defaults increase a little, the losses could be low because a bank can foreclose on a home and sell it at a reasonable price. Investor losses on 2001–2004 loans could have been low enough to allow them to ignore any warning signs. By the end of 2008, after the surge in defaults for 2005–2006 loans, activity in the subprime loan market declined dramatically, with an abrupt decrease in securitization.

Another facet of securitization being a repeated game is that the desire to build and maintain a reputation could provide an issuer with an incentive to collect soft

[24] As another example, once loan defaults had increased in the third quarter of 2007, in November 2007 Standard and Poor's adjusted its default model to reduce the reliance on the FICO score as a predictor of default (Standard and Poor's, 2007).

information and preserve its loan quality over time. In a theoretical model, Mathis, McAndrews, and Rochet (2009) consider whether reputation concerns can discipline rating agencies, and find that the answer is affirmative only if the fraction of income earned by rating agencies from rating complex products is sufficiently low. Applied to issuers, their model implies that lenders with a large proportion of income coming from securitized subprime loans would be especially vulnerable to the effects we find.

9. Conclusion

Establishing a liquid market for a complicated security requires standardization of not just the terms of the security, but also of the fundamental valuation model for the security, both of which help investors to better understand the security. Inevitably, the process of constructing and validating a model includes testing it against previous data. We argue in this paper that the growth of the secondary market for a security can have an important incentive effect that affects the quality of the collateral behind the security itself. The associated regime change implies that even a model that fits historical data well necessarily fails to predict cash flows, and hence values, going forward.

While we focus on a particular statistical default model, similar models are widely used by market participants for diverse purposes such as making loans to consumers (for example, using the FICO score), assessing capital requirements on lenders, and determining the ratings of CDO tranches. Our critique applies to all such models, because they all use historical data in some manner to predict future defaults without accounting for the impact of changed incentives of participants that generate the data. Importantly, the effects we find are systematic and stronger for borrowers with low FICO scores and low documentation. Because the loans we analyze represent the underlying collateral for CDOs and subsequent securitization, the errors cannot be diversified away. The phenomenon we examine is therefore different from the much-discussed argument that correlations (but not levels) of loan defaults had been misestimated.

The inescapable conclusion of a Lucas critique is that actions of market participants could undermine any rigid regulation. What can agents do to better predict the future? Regulators setting capital requirements or rating agencies take some time to learn about the exact magnitudes of relevant variables following a regime change. Nevertheless, we certainly expect them to be aware that incentive effects could lead to such a regime change, which can systematically bias default predictions downward. An adaptive learning approach that places more weight on recent data could help in such a setting. Once sufficient data have accumulated in the new regime, a statistical model can be reliably estimated (until the regime changes yet again). During the learning phase, however, participants need to be particularly aware that predictions from the default model are probabilistic and the set of possible future scenarios has expanded in an adverse way. Thus, the assessment of default risk must be extra conservative during this period.

We expect that the agents in the market eventually learn that the regime has changed. The challenge for regulators in particular is to recognize such shifts in real time and take appropriate actions. If investors are rational, market prices should reflect the risk of assets and could be used by regulators to assess default risk. Another alternative is to use a structural approach. In the regulatory context, perhaps a regulator can require greater disclosure of data collected by a lender, even if not reported to an investor. Such data can then be used in a structural framework to properly determine the default risk of loans by accounting for changes in the behavior of agents in response to a change in incentives (for example, by augmenting the statistical default model with a selection equation, as highlighted in the Appendix).

Appendix A. Selection model

In this Appendix, we use the selection model framework of Heckman (1980) to discuss our hypothesis that the mapping between observables and loan defaults will change with securitization. Recall that X_{it} consists of variables reported by the lender to the investor and Z_{it} of variables observed by the lender but not reported to the investor. For convenience, assume that X_{it} and Z_{it} are both non-negative scalars, denoted respectively by x_{it} and z_{it}. For example, x_{it} could be the FICO score of the borrower and z_{it} could be a summary statistic based on other hard and soft information available to the lender.

A regulator or rating agency has the same information as the investor, and is interested in evaluating the quality of the loan based on x_{it}. Let d_{it} represent a default event on loan i issued at time t. A contemporaneous default regression could be estimated as

$$d_{it} = \alpha + \beta x_{it} + \epsilon_{it}, \tag{11}$$

where ϵ_{it} is a mean zero error term with variance σ_ϵ^2.[25]

In a low-securitization regime, the lender approves a loan application if either x_{it} is high or x_{it} is low but z_{it} is high. That is, $A_{it} = 1$ if and only if $\gamma z_{it} + \delta x_{it} + \eta_{it} > 0$, where η_{it} is a mean-zero error term with variance σ_η^2. The regulator, rating agencies and the investors observe only approved loans (i.e., $A_{it} = 1$).

Assume that the conditional expectation of ϵ_{it} given η_{it} is linear in η_{it} and that the correlation between ϵ_{it} and η_{it} is ρ. Then, we can write $\epsilon_{it} = \rho(\eta_i - \overline{\eta})\sigma_\epsilon/\sigma_\eta + \omega_{it}$, where ω_{it} is uncorrelated with η_{it}. Therefore, $\mathrm{E}(d_{it}|x_{it}, A_{it} = 1) = \beta x_{it} + (\rho\sigma_\epsilon/\sigma_\eta)\mathrm{E}(\eta_{it}|\eta_i > -\gamma z_{it} - \delta x_{it})$.

In the spirit of Olsen (1980), assume that η_{it} is uniformly distributed over $[-1, 1]$. Then, $\mathrm{E}(\eta_{it}|\eta_i > -\gamma z_{it} - \delta x_{it}) = (1 - \gamma z_{it} - \delta x_{it})/2$. It follows that

$$\mathrm{E}(d_{it}|x_{it}, A_{it} = 1) = \beta x_{it} + \frac{\rho\sigma_\epsilon}{2\sigma_\eta}\left[-\delta x_{it} - \gamma z_{it} + 1\right]. \tag{12}$$

Therefore, when Eq. (11) is estimated, the relation between the observed coefficient β^* and the true coefficient

[25] Although default is a binary event, here we use a linear regression specification for expositional simplicity. The analysis is similar with a logit or probit specification. Our actual regressions in Section 5 use the logit model.

β can be written as $\beta^* = \beta + (\rho\sigma_\epsilon/2\sigma_\eta)[-\delta Var(x_{it}|A_{it}=1) - \gamma Cov(x_{it}, z_{it}|A_{it}=1)]$. Here, $Var(x_{it}|A_{it}=1) > 0$. Further, the selection equation implies on average that, for high values of x_{it}, $A_{it}=1$ even when z_{it} is low. However, for low values of x_{it}, on average $A_{it}=1$ only when z_{it} is high. Thus, $Cov(x_{it}, z_{it}|A_{it}=1) < 0$. Let $B_\ell = \beta - \beta^* = (\rho\sigma_\epsilon/2\sigma_\eta)[\delta Var(x_{it}|A_{it}=1) + \gamma Cov(x_{it}, z_{it}|A_{it}=1)]$ denote the bias in the low-securitization regime.

Next, consider a high securitization regime. Here, the lender bases its decisions on hard information variables that are reported to the investor, downplaying information it could have used in a low-securitization regime. In the extreme case, if z_{it} is completely ignored, the selection equation changes to $A_{it}=1$ if and only if $\delta_h x_{it} + \eta_{it} > 0$, where δ_h is sufficiently greater than δ to ensure that the minimum value of x_{it} at which a loan is granted is the same in both regimes. That is, even when x_{it} is small, on average the loan is granted regardless of the value of z_{it}. Here, $Cov(x_{it}, z_{it}|A_{it}=1) = 0$. Therefore, the bias in the high-securitization regime could be represented as $B_h = (\rho\sigma_\epsilon/2\sigma_\eta)\delta_h Var(x_{it}|A_{it}=1)$, where we assume that $Var(x_{it}|A_{it}=1)$ is similar in both regimes.

Because the true coefficient β is negative (that is, when the FICO score x_{it} is high, a default is less likely), the estimated coefficient in the low-securitization regime (say β^*_ℓ) is closer to zero due to additional covariance term than the coefficient in the high-securitization regime (β^*_h). Therefore, if β^*_ℓ is used to forecast defaults for low values of x_{it}, it underestimates defaults.[26] Because defaults themselves are more likely at low values of x_{it}, the overall effect is to underpredict defaults in the high-securitization era.

Overall, then, our argument is that regulators, rating agencies, and investors see only approved loans, which by definition have survived a selection process. The selection process for loans changes when the incentives of the lender change. Consequently, as securitization increases, one expects that the behavior of the lender changes. This changes the selection process, thereby altering the mapping from observables to loan defaults.

References

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

Agarwal, S., Hauswald, R., 2010. Distance and private information in lending. Review of Financial Studies 23, 2757–2788.

Ashcraft, A., Goldsmith-Pinkham, P., Vickery, J., 2010. MBS ratings and the mortgage credit boom. Staff Report No. 449, Federal Reserve Bank of New York, NY.

Basel Committee on Banking Supervision, 2006. International convergence of capital measurement and capital standards: a revised framework, comprehensive version ⟨http://www.bis.org/publ/bcbs128.pdf⟩.

Benmelech, E., Dlugosz, J., 2009. The alchemy of CDO credit ratings. Journal of Monetary Economics 56, 617–634.

Bolton, P., Faure-Grimaud, A., 2010. Satisficing contracts. Review of Economic Studies 77, 937–971.

Brunnermeier, M., 2009. Deciphering the liquidity and credit crunch 2007–2008. Journal of Economic Perspectives 23, 77–100.

Brunnermeier, M., Crockett, A., Goodhart, C., Persaud, A., Shin, H., 2009. Geneva Reports on the World Economy 11: The Fundamental Principles of Financial Regulation. Centre for Economic Policy Research, London, England.

Calomiris, C., 2009. The debasement of ratings: What's wrong and how we can fix it. Unpublished working paper. Columbia University, New York, NY.

Cole, R., Goldberg, L., White, L., 2004. Cookie-cutter versus character: the micro structure of small business lending by large and small banks. Journal of Financial and Quantitative Analysis 39, 227–251.

Coval, J., Jurek, J., Stafford, E., 2009. Economic catastrophe bonds. American Economic Review 99, 628–666.

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

Faltin-Traeger, O., Johnson, K., Mayer, C., 2010. Issuer credit quality and the price of asset-backed securities. American Economic Review 100, 501–505.

Fishelson-Holstein, H., 2005. Credit scoring role in increasing homeownership for underserved populations. In: Retsinas, P., Belsky, E. (Eds.), Building Assets, Building Credit: Creating Wealth in Low-Income Communities, Brookings Institution Press, Washington, DC.

Gerardi, K., Lehnert, A., Sherlund, S., Willen, P., 2008. Making sense of the subprime crisis. Brookings Papers on Economic Activity 39, 69–159.

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

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

Greenspan, A., 2008. Testimony before House Committee of Government Oversight and Reform. October 23.

Griffin, J., Tang, D., 2012. Did subjectivity play a role in CDO credit ratings? Journal of Finance 67, 1293–1328.

Heckman, J., 1980. Varieties of selection bias. American Economic Review 80, 313–318.

Holloway, T., MacDonald, G., Straka, J., 1993. Credit scores, early-payment mortgage defaults, and mortgage loan performance. Unpublished working paper. Federal Home Loan Mortgage Corporation, Tysons Corner, VA.

Holmström, B., Milgrom, P., 1990. Multi-task principal-agent problems: incentive contracts, asset ownership and job design. Journal of Law, Economics, and Organization 7, 24–52. (special issue).

Inderst, R., Ottaviani, M., 2009. Misselling through agents. American Economic Review 99, 883–908.

Jiang, W., Nelson, A., Vytlacil, E., 2014. Securitization and loan performance: a contrast of ex ante and ex post relations in the mortgage market. Review of Financial Studies 27 (2), 454–483.

Kashyap, A., Rajan, R., Stein, J., 2008. Rethinking capital regulation. Paper prepared for the Federal Reserve Bank of Kansas City Symposium, Jackson Hole, WY.

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

Keys, B., Seru, A., Vig, V., 2012. Lender screening and the role of securitization: evidence from prime and subprime mortgage markets. Review of Financial Studies 25, 2071–2108.

Kroszner, R., 2007. Innovation, information, and regulation in financial markets. Speech at the Philadelphia Fed Policy Forum, Philadelphia, PA, November 30, 2007.

Liberti, J., Mian, A., 2009. Estimating the effect of hierarchies on information use. Review of Financial Studies 22, 4057–4090.

Loutskina, E., Strahan, P., 2011. Informed and uninformed investment in housing: the downside of diversification. Review of Financial Studies 24, 1447–1480.

Lucas, R., 1976. Econometric policy evaluation: a critique. In: Brunner, K., Meltzer, A. (Eds.), The Phillips Curve and Labor Markets Carnegie-Rochester Conferences on Public Policy, North Holland Press, Amsterdam, pp. 19–46.

Mathis, J., McAndrews, J., Rochet, J.-C., 2009. Rating the raters: are reputation concerns powerful enough to discipline rating agencies?. Journal of Monetary Economics 56, 657–674.

Mayer, C., 2010. Housing, subprime mortgages, and securitization: how did we go wrong and what can we learn so this doesn't happen again? Testimony before Financial Crisis Inquiry Commission ⟨http://fcic.law.stanford.edu⟩.

Mayer, C., Pence, K., 2008. Subprime mortgages: what, where, and to whom? Unpublished working paper no. 14083. National Bureau of Economic Research, Cambridge, MA.

Mayer, C., Pence, K., Sherlund, S., 2009. The rise in mortgage defaults. Journal of Economic Perspectives 23, 27–50.

[26] In other words, the bias with respect to the true coefficient changes across the two regimes. In particular, because $Cov(x_{it}, z_{it}|A_{it}=1) < 0$ in the low-securitization regime and $\delta_h > \delta$, it follows that $B_h > B_\ell$.

Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the US mortgage default crisis. Quarterly Journal of Economics 124, 1449–1496.
Mian, A., Sufi, A., Trebbi, F., 2011. Foreclosures, house prices, and the real economy. Unpublished working paper no. 16685. National Bureau of Economic Research, Cambridge, MA.
Norris, F., 2006. 30-Year Treasury Bond returns and demand is strong. New York Times, February 9.
Olsen, R., 1980. A least squares correction for selectivity bias. Econometrica 48, 1815–1820.
Pagano, M., Volpin, P., 2010. Credit ratings failures and policy options. Economic Policy 25, 401–431.
Pence, K., 2006. Foreclosing on opportunity? State laws and mortgage credit. Review of Economics and Statistics 88, 177–182.
Piskorski, T., Seru, A., Vig, V., 2010. Securitization and distressed loan renegotiation: evidence from the subprime mortgage crisis. Journal of Financial Economics 97, 369–397.
Rajan, U., Seru, A., Vig, V., 2010. Statistical default models and incentives. American Economic Review 100, 506–510.
Sheather, S., Jones, M., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B 53, 683–690.
Standard and Poor's, 2007. Standard & Poor's enhances LEVELS® 6.1 model. News release, November 9, 2007 ⟨www2.standardandpoors.com⟩.
Stein, J., 2002. Information production and capital allocation: decentralized versus hierarchical firms. Journal of Finance 57, 1891–1921.
Tirole, J., 2009. Cognition and incomplete contracts. American Economic Review 99, 265–294.