Predicting loss severities for residential mortgage loans: A three-step selection approach[*]

Hung Xuan Do[a,b], Daniel Rösch[c], Harald Scheule[a†]

[a]*Finance Discipline Group, UTS Business School, University of Technology, Sydney, Australia*

[b]*School of Economics and Finance, Massey University, Albany campus, Auckland, New Zealand*

[b]*Department of Statistics and Risk Management, Faculty of Economics, University of Regensburg, Germany*

This version: 28 July 2017

[*] The authors would like to thank the participants of the seminars at the University of Regensburg, Germany, and the University of Technology, Sydney. We also gladly acknowledge the valuable research assistance of Michael Tjendara. The support by the Center for International Finance and Regulation (CIFR, project number E001) is gratefully acknowledged. CIFR is funded by the Commonwealth and NSW Governments and supported by the other consortium members (see www.cifr.edu.au). Hung Do acknowledges financial research support from the Finance Discipline Group, UTS Business School.

[†] Corresponding author. Tel. +61 2 9514 7724.

Email addresses: h.do@massey.ac.nz (Hung Do), daniel.roesch@ur.de (Daniel Rösch), harald.scheule@uts.edu.au (Harald Scheule).

Predicting loss severities for residential mortgage loans: A three-step selection approach

ABSTRACT

This paper develops a novel framework to model the loss given default (LGD) of residential mortgage loans which is the dominant consumer loan category for many commercial banks. LGDs in mortgage lending are subject to two selection processes: default and cure, where the collateral value exceeds the outstanding loan amount. We propose a three-step selection approach with a joint probability framework for default, cure (i.e., zero-LGD) and non-zero loss severity information. The proposed methodology dominates widely used ordinary least squares regressions for LGDs in terms of out-of-time predictions.

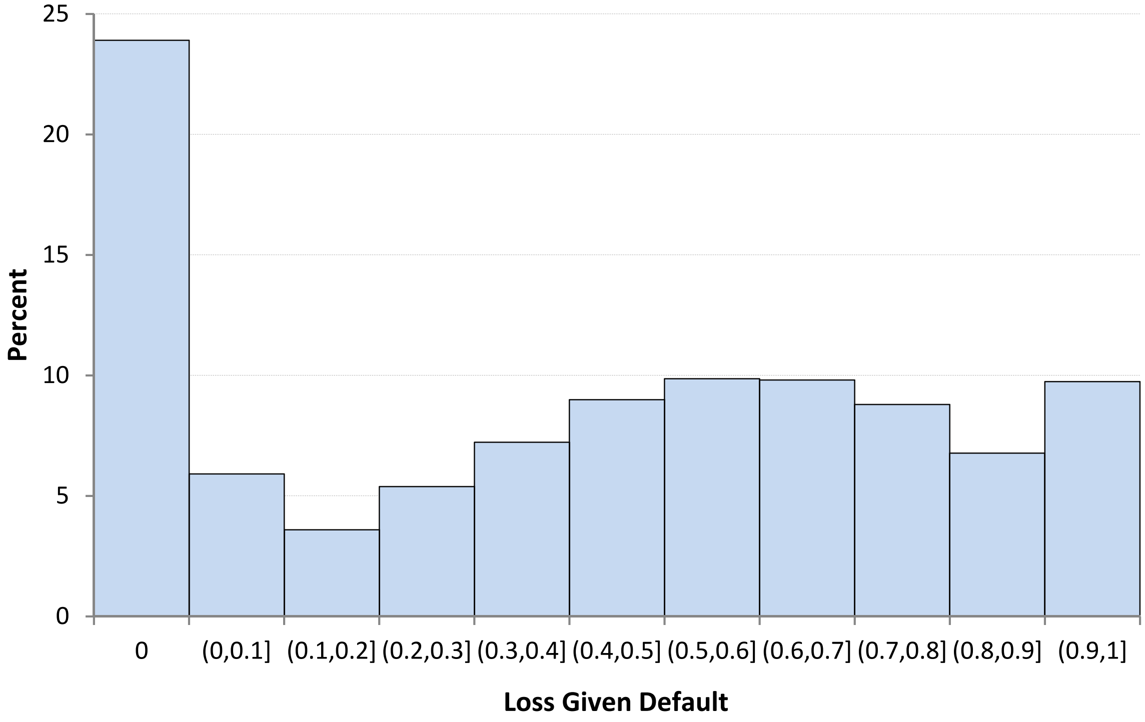
Keywords: Analytics, Default, Loss given default, Residential Mortgage, Selection Model.

JEL classification: G21, G28, C19.

1. Motivation and literature review

Residential mortgage loans are key assets on bank balance sheets. The International Monetary Fund reports under its financial soundness indicators real-estate loan to total loan exposures of banks of 63.5% for Australia, 18.5% for Germany, 36.6% for Canada and 31.0% for the US. Note that these numbers are conservative as they do not include non-bank lenders and loan portfolios sold to non-banks via securitisation and other structures.

Figure 1: Distribution of Loss Given Default



With regard to loss rate given default (LGD, hereafter) modelling selection models have been proposed. First, loss rates given default can only be observed if a default event occurs. Bade et al. (2011) and Roesch and Scheule (2014) model US corporate LGD conditional on the selecting default process. Andersson and Mayock (2014) model Florida mortgage LGDs conditional on the selecting default process. Second, a large fraction of defaulted loans does not generate losses. Previous studies on personal loans (e.g., credit card loans) have highlighted a necessity to treat zero and non-zero losses differently (see Matuszyk et al., 2010; Loterman et al., 2012, Bijak and Thomas, 2015; and Yao et al., 2017). Mortgage loans are collateralised by the funded houses and hence loss risk is closely related to house prices. Given default, house prices often exceed the linked loan exposure and mortgages cure, i.e., result in zero or low

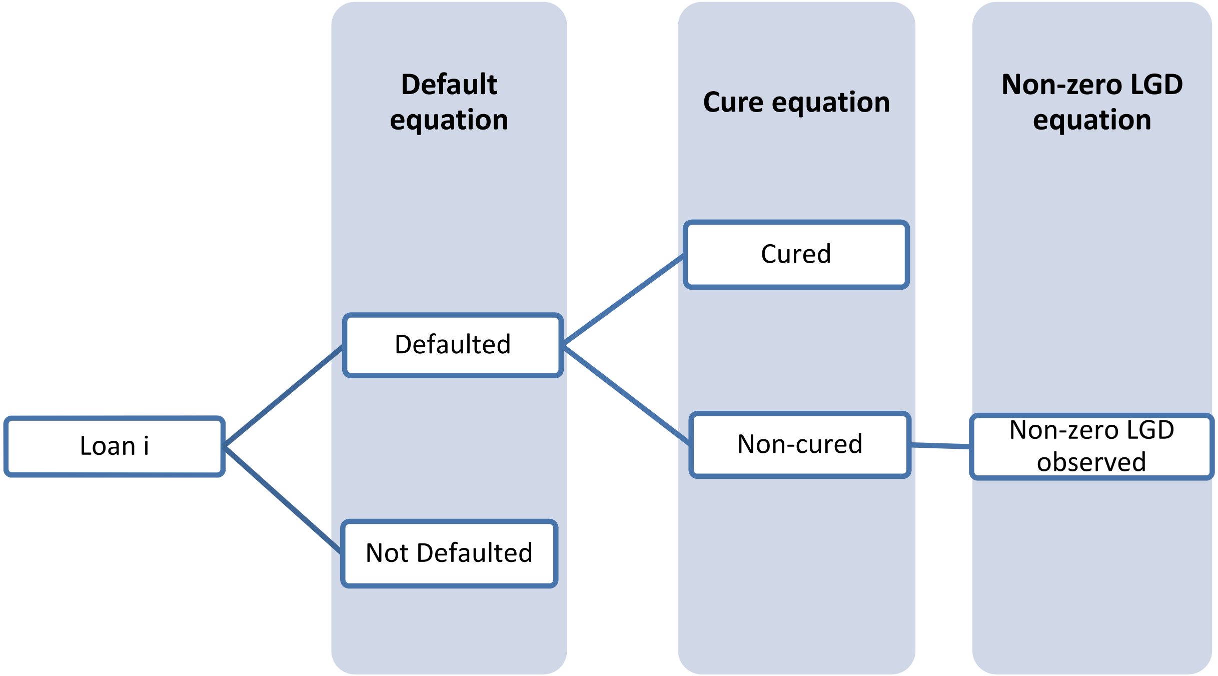
LGDs. Figure 1 shows that the empirical LGD distribution for mortgage LGDs has a bimodal characteristic with a very high peak at zero.

These features are not reflected in contemporary industry risk models. Commercial banks estimate probabilities of default (PD, hereafter) and LGDs for loan loss provisioning, regulatory capital requirements and loan pricing. These estimation models mainly apply non-linear regression techniques (e.g., Probit models) for PDs and linear regression techniques (e.g., ordinary least squares regressions) for LGDs.

As a response to this need, in this paper we are first to model residential mortgage LGDs conditional on both the default and cure process. Figure 2 illustrates this approach which estimates the three stages jointly.

Figure 2: Selection mechanism among default, cure and non-zero LGD

This figure shows the selection mechanism among default, cure and non-zero LGD that our model is based on. The mechanism indicates that the cure events are observed if loan *i* defaults, while the non-zero LGD can only be observed if the defaulted loan *i* is non-cured. This mechanism requires a joint probability framework between default, cure and non-zero losses for modelling and prediction purpose.



This paper develops a methodology to jointly model defaulted, cured and non-cured loans in three stages: the first stage models the probability of default, which is important for avoiding a sample-section bias, the second stage models the probability of cure and the third stage models the magnitude of non-zero LGD. The contributions of our paper are as follows, First, we provide a novel methodology that provides a joint forecasting model for default probabilities and loss rates given default. Second, the methodology controls for the selection bias inherent in the default process as well as the non-cure process by simultaneously estimating all three equations. Third, we measure the correlation between the default, cure and loss processes which is an important constituent in portfolio risk modelling for bank capital. Fourth, we provide empirical evidence on time-varying risk factors for a large dataset on over 2.5 million US single family mortgage loans with over 29 million quarterly observations and 585,781 default events originated from 1990 and observed between 2004 and 2015. Specifically, our study contributes two new drivers of the credit risk quantities, namely the average local loan foreclosure rate and the average gap between local house price under a common market condition and under a distress condition. Fifth, we analyse the econometric merits of the new methodology in a comparison with standard industry models. This methodology dominates common ordinary least squares (OLS) regressions by providing better predictions for future LGDs.

The paper relates to the literature on consumer PD and LGD estimation. With regard to PDs, the literature has generally focused on the credit card and other personal loans with various modelling approaches. PDs are modelled using probit and logistic regression models (e.g., Crook et al., 2007; Crook and Bellotti, 2010; Leow and Mues, 2012; Lee et al., 2016), survival analyses (e.g., Belloti and Crook, 2008; Malik and Thomas, 2009; Tong et al., 2012), multiple classifiers (e.g., Finlay, 2011; Zhang et al., 2014), neural networks (e.g, Baesens et al., 2003; Akkoc, 2012) and support vector machines (e.g., Maldonado et al., 2017). In terms of residential mortgage loans, previous studies have employed various borrower, loan and macroeconomic characteristics as risk drivers. For instance, Amromin and Paulson (2009) document that real estate price is an important risk driver. Elul et al. (2010) show an important role of borrowers' credit scores at origination (i.e., FICO scores) on predicting mortgage PDs. Crook and Banasik (2012) exploit mortgage rate and real house price index to predict mortgage default rates. Rajan et al. (2015) indicate that change in lender incentives during the securitization boom exacerbated an accuracy of the statistical mortgage default predictive model.

With regard to mortgage LGDs, current practice and literature have widely employed OLS regression for modelling purposes yet various determinants of loss severity have been examined. For example, a number of loan characteristics including loan-to-value (LTV), loan age, loan size, loan type, and loan purpose have been investigated (see, Clauretie and Herzog, 1990; Lekkas et al., 1993; Crawford and Rosenblatt, 1995; Pennington-Cross, 2003; Calem and LaCour-Little, 2004; Qi and Yang, 2009; Zhang et al., 2010; Leow and Mues, 2012 and Johnston Ross and Shibut, 2015). Property characteristics including owner occupancy, property types, (e.g., single family, condominium and manufactured house) and state foreclosure laws (judicial process, statutory right of redemption, deficiency judgment) are examined in Clauretie and Herzog (1990), Crawford and Rosenblatt (1995), Pennington-Cross (2003), Qi and Yang (2009), Zhang et al. (2010) and Johnston Ross and Shibut (2015). Economic and market conditions including housing market conditions, change in unemployment rate, real economic growth and median income are considered in Clauretie and Herzog (1990), Calem and LaCour-Little (2004), Qi and Yang (2009), Zhang et al. (2010) and Leow et al. (2014).

Another literature stream analyses the link between PD and LGD. The majority of these studies relate to unsecured credit exposures such as corporate and credit card loans. Examples are Altman et al. (2005), Chava et al. (2011), Bade et al. (2011), Bellotti and Crook (2012) and Roesch and Scheule (2014).

The remainders are organised as follows. Section 2 describes the data source, the definitions of default, cure and non-zero LGD as well as explanatory variables used in analysis. Section 3 develops an econometrics framework for modelling LGDs which follows three steps of selection mechanism from default to non-zero loss severity. Section 4 discusses empirical results and section 5 concludes the paper.

2. Data

Our sample is provided by the International Financial Research database, which collects monthly loan level data for U.S. non-agency residential mortgage-backed securities. We filter for first lien loans for single-family[3] observed at quarterly intervals from Q1 2004 to Q1 2015, which comprises 2,531,346 loans and produces a total of 29,032,350 observations. The complete dataset contains information on the loan (e.g., actual loan balance at origination and observation time, scheduled loan balance at the end of each period, loan age, interest rate and loan type), on the property (e.g., property location, property value at origination, whether the property is occupied by owner), and on the borrower (e.g., FICO score at origination). Further, the data includes information about the modifications and foreclosure as well as losses on liquidated property and losses on previously liquidated loans resulted from the foreclosure[4].

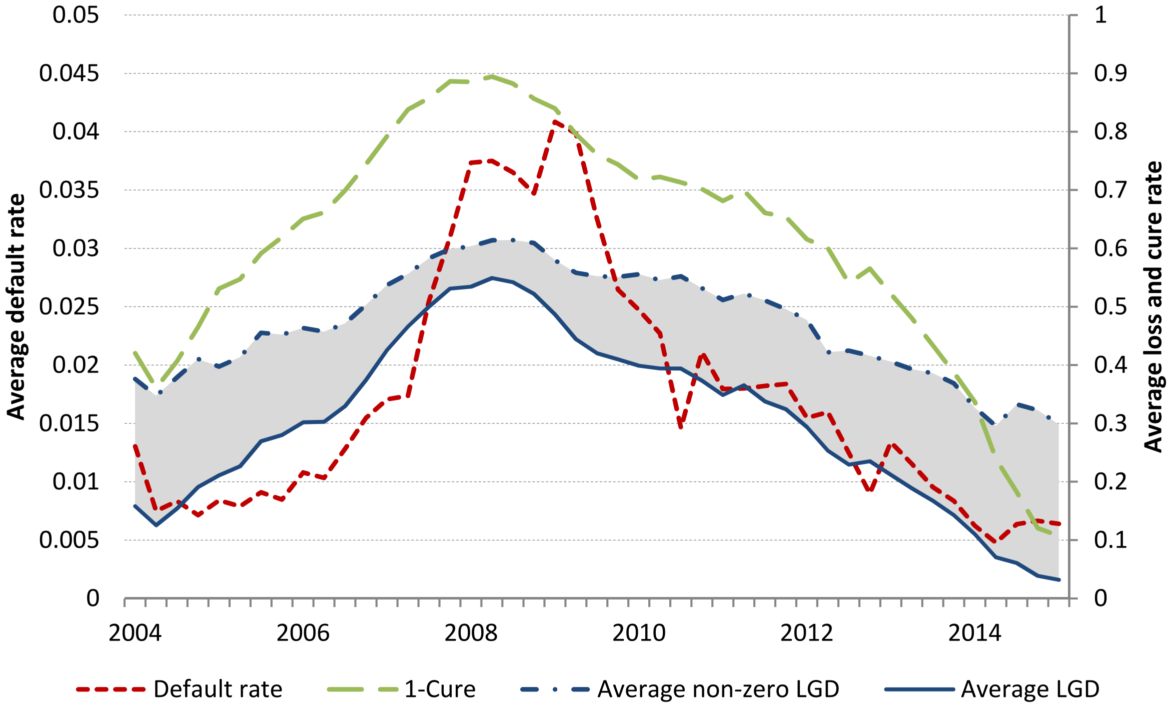
[3] We focus on the first lien loan for single-family due to our access to data of the house price index at the MSA level is only available for single-family home. The house price index is a critical variable in our analysis since it is used to calculate the Current Loan-to-Value ratio.

[4] Normally, information about gains or losses is recorded monthly in loan level reports. There are two separate fields: (i) the original loss is recorded in a field called "Loss on Liquidated Property" and (ii) subsequent losses are recorded in a field called "Loss on previous Liquidated Loans". Hence, for a precise measure of loss, we employ information in both of these fields.

2.1 Default, cure and non-zero loss given default

We define the default event as loan foreclosure. We observe 585,781 defaulted single-family mortgage loans between 2004 and 2015, which peaked during the Global Financial Crisis (GFC, see Figure 3).

Figure 3: Default rate, cure rate and Loss Given Default by time



We define the LGD of a loan as the sum of discounted losses on liquidated property and previous liquidated loans, divided by the actual loan balance at default. We follow Qi and Yang (2009) to discount losses to the time of default using 1-year LIBOR plus 3%[5]. Hence, the LGD of loan *i* defaulted at time t_d can be calculated as:

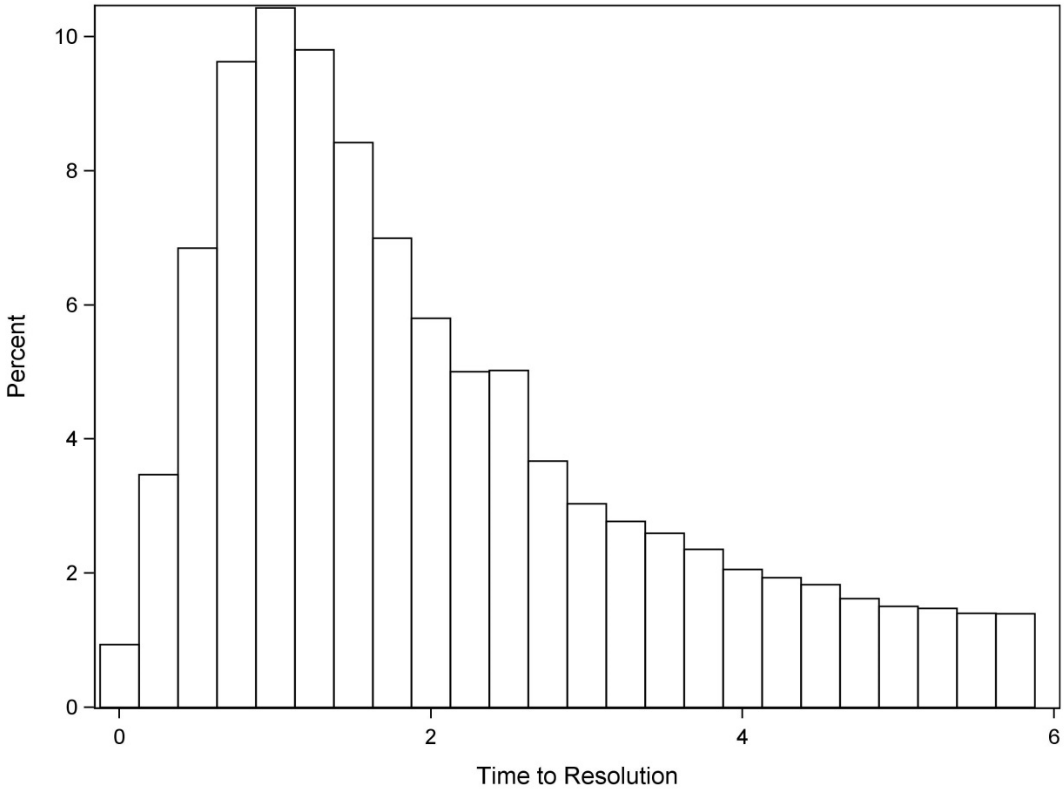
$$LGD_{i,t_d} = \frac{1}{Outstanding\ Balance_{i,t_d}} \times \sum_{t=t_d}^{T} \frac{Losses_{i,t}}{(1+r)^{t-t_d}} \qquad (1)$$

where *r* is the 1-year LIBOR plus 3%. The discount rate is consistent with common discount rate models (see Jortzik and Scheule, 2017). Our data contains information about the losses resulted by loan defaults until Quarter 1 of 2015. As can be seen from Figure 4, most of non-zero losses arose within 3 years after the default events. However, there is also a possibility that the losses will arise several years after the time of default. Due to the accounting

[5] It is worth noting that this is equivalent to a weighted average cost of capital concept in which a bank allocates 50% of capital to the defaulted exposure and a market risk premium of 6%. We consider this number to be very conservative due to the fact that mortgage exposures are collateralised and the systematic variation of LGDs substantially lower than for unsecured exposures

terminology losses build up over time (which is different to accounting systems where the outstanding loan amount is recorded at default and recovery cash flows are recorded over time). We control for a possible bias that losses are understated as not all losses are realised (resolution bias) by the variable TimeToEOO, which measures the time gap between default events and the time that the last loss was observed.

Figure 4: Distribution of time after default that non-zero LGD observed



As calculated in Eq. (1), we define cure and non-cure as zero and non-zero LGD, respectively. We summarise the default rate, cure rate and average of LGD per origination and observation year in Table 1. Consistent with Demyanyk and Hemert (2011), the majority of default loans are originated in years immediately prior to the GFC or observed in years during the GFC. Further, both average LGD (including zero-LGD) and average non-zero LGD (excluding non-zero LGD) in the origination years immediately before the GFC and observation years during the GFC are the highest compared to other years, which vary between 40% and 55%. However, this behaviour is opposite to the cure rate, which experiences troughs (equivalently, peaks with one minus the cure rate) during these years. The reverse movements of these variables can be seen in more detail from Figure 3, which shows that since 2004 the trend of non-zero LGDs is quite in line with the default rate but completely inverse with the cure rate.

Table 1: Default rates, cure rates and average LGD per origination and observation years

This table reports the default rates (default per number of loans, **D/N**), cure rates (cure per number of defaults, **C/D**), average LGD ($\overline{\boldsymbol{LGD}}$**,** i.e., average of both zero and non-zero LGD) and average non-zero LGD ($\overline{\boldsymbol{LGD}}^*$) according to Origination Years and Observation Years.

Origination Years								**Observation Years**							
Year	**N**	**D**	**D/N**	**C**	**C/D**	$\overline{LGD}$	$\overline{LGD}^*$	**Year**	**N**	**D**	**D/N**	**C**	**C/D**	$\overline{LGD}$	$\overline{LGD}^*$
1990	127,962	862	0.007	592	0.687	0.120	0.384								
1991	3,230	38	0.012	26	0.684	0.054	0.170								
1992	6,052	51	0.008	35	0.686	0.079	0.252								
1993	15,430	134	0.009	99	0.739	0.063	0.241								
1994	18,627	165	0.009	104	0.630	0.128	0.346								
1995	28,009	305	0.011	183	0.600	0.219	0.548								
1996	31,066	367	0.012	222	0.605	0.220	0.556								
1997	63,905	855	0.013	464	0.543	0.263	0.575								
1998	113,325	1,598	0.014	846	0.529	0.269	0.572								
1999	134,415	2,668	0.020	1,173	0.440	0.339	0.605								
2000	133,568	2,861	0.021	1,388	0.485	0.315	0.613								
2001	195,505	3,760	0.019	1,601	0.426	0.284	0.494								
2002	815,093	12,448	0.015	5,285	0.425	0.266	0.463								
2003	3,015,024	23,569	0.008	10,086	0.428	0.261	0.457								
2004	5,542,684	60,322	0.011	20,646	0.342	0.310	0.471	**2004**	968,012	8,102	0.008	4,697	0.580	0.162	0.385
2005	7,808,216	140,291	0.018	32,491	0.232	0.404	0.526	**2005**	2,381,583	20,240	0.008	8,538	0.422	0.252	0.435
2006	8,407,766	256,998	0.031	47,255	0.184	0.479	0.587	**2006**	3,757,697	47,031	0.013	14,187	0.302	0.335	0.479
2007	2,335,650	74,395	0.032	16,134	0.217	0.466	0.595	**2007**	4,748,736	108,707	0.023	16,040	0.148	0.491	0.576
2008	196,653	3,940	0.020	1,319	0.335	0.334	0.503	**2008**	4,098,517	150,027	0.037	17,837	0.119	0.538	0.610
2009	4,465	70	0.016	32	0.457	0.406	0.748	**2009**	3,098,650	109,693	0.035	22,612	0.206	0.447	0.563
2010	1,784	11	0.006	4	0.364	0.305	0.480	**2010**	2,412,271	50,476	0.021	14,401	0.285	0.391	0.547
2011	6,486	4	0.001	3	0.750	0.240	0.959	**2011**	2,001,138	36,293	0.018	11,826	0.326	0.344	0.511
2012	27,435	69	0.003	34	0.493	0.137	0.270	**2012**	1,770,128	23,618	0.013	9,782	0.414	0.258	0.440
								2013	1,814,842	19,553	0.011	10,430	0.533	0.183	0.393
								2014	1,606,328	9,643	0.006	7,531	0.781	0.070	0.321
								2015	374,448	2,398	0.006	2,141	0.893	0.032	0.299
Total	**29,032,350**	**585,781**	**0.020**	**140,022**	**0.239**				**29,032,350**	**585,781**	**0.020**	**140,022**	**0.239**		

The observed behaviours of these variables are consistent with the fact that a large number of defaults and residential mortgage loan losses were experienced during the GFC. These movements suggest a positive correlation structure between default rates and non-zero LGDs at *their mean level* yet a negative correlation structure between them and the cure rate. These results share the same intuition with previous theories and empirical evidence in the corporate loan literature: if a loan defaults, the LGD may depend on the value of loan collateral. Similar to the value of other assets, the value of collateral is subject to changes in the economic environment. In case if the economy experiences a downturn, the LGD may increase just as the likelihood of default tends to increase, and, therefore, a positive association between LGD and PD in *their mean level* can be observed (see Frye, 2000a, 2000b; Jarrow, 2001; Jokivuolle and Peura, 2003; and Altman et al., 2005, among others). Likewise, this explanation can also be extended to support the negative correlation between PD (or non-zero LGD) and the probability of cure (PC) in *their mean level*.

Comparing the average LGD and average non-zero LGD, we observe a time varying gap between them (the grey area in Figure 3), which is mainly due to the changing behaviour of cure rate. This indicates different dynamics between cured and non-cured loans and, hence, motivate us to model zero and non-zero LGDs in separate selection steps. Further, it is also worth noting that the average loss severity approaches zero at the end of our observation time (Q1 2015), which is a consequence of the aforementioned resolution bias. Andersson and Mayock (2014) suggest including the resolution period in modelling LGD to overcome this possible loss information bias. However, banks are often unable to determine resolution if they continue to hold a claim against the borrower. We control for this potential bias by utilising a variable that captures an effect of time gap between default events and the time of last available loss information (i.e., Q1 2015).

We depict the distribution of the LGD in Figure 1, which shows the largest proportion (around 30%) of LGD observations is the [0.0, 0.1) range, in which around 25% of the total LGDs observations are zero-LGDs (or cures). This is consistent with our expectation and the literature (e.g., Leow and Mues, 2012; and Tong et al., 2013). However, the distribution of non-zero LGD in our data sample is different to Leow and Mues (2012) and Tong et al. (2013). The two mentioned studies used the UK residential mortgage loans that defaulted before 2001 and found that the distribution of nonzero-LGD concentrates to levels considerably less than 50% and very few observed LGD are greater than 50%. Since our data also covers the US residential mortgage loans that defaulted from 2004 and especially during the GFC, a much greater proportion of non-zero LGD observations located in the far right-hand side of the distribution are expected.

2.2 Explanatory variables

We employ various variables to explain for the dynamics of the LGD, probability of default and probability of cure, including loan characteristics, borrower characteristics, property characteristics and economic and market conditions. Among all explanatory variables under consideration, current loan-to-value ratio (*CLTV*) is one of the most important drivers of credit risk quantities. We calculate the *CLTV* using the House Price Index (HPI) at Metropolitan Statistical Areas (MSAs) level collected from Federal Housing Finance Agency (FHFA). We first map the zip-code of the property to the MSA of the HPI by using mapping data provided by the US Department of Housing and Urban Development. We exclude loans with missing zip-codes. If a zip-code can map to more than one MSA, we choose the MSA that has a dominant residential ratio (mostly greater than 90%, which means more than 90% of residents live in that MSA). The *CLTV* of loan i at observation time t is calculated as, $CLTV_{i,t} = \frac{CB_{i,t}}{CAV_{i,t}}$, where $CB_{i,t}$ is the current balance (or outstanding balance) of loan i observed at time t. The current appraisal value (*CAV*) is approximated by the following formula, $CAV_{i,t} = OAV_i \times \frac{HPI_{i,t}}{HPI_{i,t_0}}$, where *OAV* is the appraisal value at origination time and t_0 indicate the origination time.

While other explanatory variables have been documented in the literature, our study contributes two new drivers of the credit risk quantities, namely Average local loan foreclosure rate (*FCRate*) and Average gap between local house price under a common market condition and under distress condition (*GapHprice*).

The *FCRate* is calculated as:

$$FCRate_{r,t} = \frac{N_{d,r,t}}{N_{l,r,t}} \quad (2)$$

where r denotes the region (i.e., MSA level in our analysis). $N_{d,r,t}$ denotes number of default loans in the region r at time t. $N_{l,r,t}$ denotes number of home loans with collaterals located in the region r at time t. Similar to zip-code – MSA mapping, we map zip-code to county by using mapping data provided by the U.S. Department of Housing and Urban Development. If a zip-code can map to more than one county, we choose the county that has a dominant residential ratio (mostly greater than 90%).

The *GapHprice* is computed as:

$$GapHprice_{r,t} = \frac{\sum_{i=1}^{N_{d,r,t}} \log\left(\frac{CAV_{i,t}}{CB_{i,t} - \sum_{j=t_d}^{T} Loss_{ij}}\right)}{N_{d,r,t}} \quad (3)$$

where, *i* indicates the loan order, t_d denotes the time of default and *T* denotes the final time that losses are realised. The losses include losses on liquidated loans and losses on liquidated property. If the outstanding loan amount ($CB_{i,t}$) or sum of losses is less than or equal to zero, $\log\left(\frac{CAV_{i,t}}{CB_{i,t} - \sum_{j=t_d}^{T} Loss_{ij}}\right)$ is assigned the value of 0. This reflects the case when the sale prices can fully cover the outstanding loan balance and associated resolution cost, which does not essentially represent the house price under distress.

Our motivation to include the *FCRate* as a driver of the credit risk measures (including PD and LGD) is from the rationale of observational learning (see Agarwal et al., 2012) and ethical/behavioural response (see Seiler et al., 2014)[6] as well as recent evidence of negative impact of local foreclosure on house prices (see for examples, Campbell et al., 2011; Anenberg and Kung, 2014; and Gerardi et al., 2015). Meanwhile, an inclusion of the *GapHprice* variable is motivated by the fact that foreclosed properties face a threat of vandalism, poor maintenance as well as protection costs, which inevitably leads to the foreclosure discounts during the repossession process.

We summarise the definition of the explanatory variables in Table 2 and provide summary statistics of variables in the three equations (including equations for PD, PC and non-zero LGDs, respectively) in Table 3.

[6] The observational learning theory affirms that borrowers tend to devaluate their property or to increasingly believe in an opinion of declining market if they observe foreclosures in their neighbourhood. Meanwhile, the ethical viewing states that an observation of many other foreclosures in the neighbourhood areas can ease the discrediting behaviour of borrowers towards default decisions.

Table 2: Definition of the explanatory variables

Variable groups	Description
Loan characteristics	
Current loan to value (*CLTV*)	$CLTV$ is calculated using the House Price Index (HPI) at the Metropolitan Statistical Areas (MSAs) level from the Federal Housing Finance Agency.
Loans size	Natural logarithm of loan amount at origination time, $Loan\ size_i = \ln(OB_i)$, where OB_i is the original balance of loan *i*.
Loan age	The number of years between the origination date and the default date. One quarter is converted to ¼ of a year.
Adjustable rate mortgage (*ARM*)	Indicator variable *ARM* represents adjustable rate mortgage (if *ARM*=1).
Modification	Indicator *Modification* receives value of 1 if a loan is modified and 0 otherwise.
Current Interest Rate (*Current IR*)	*Current IR* represents the current interest rate associated with the loan.
Borrower characteristics	
FICO	FICO represents the FICO score of a borrower at origination time, which measures the borrower's credit quality when a loan was approved.
Insufficient Loan Repayment (*ILR*)	An indicator variable represents whether an event of insufficient loan repayment is observed for loan i at time t. *ILR* variable receives value of 1 if scheduled loan balance is less than actual outstanding loan balance, indicating an insufficient loan repayment is observed.
Property characteristics	
State foreclosure laws on property	Indicator variables *NOJUD*, *SRR* and *NODJ* represent states where judicial process is not allowed (if *NOJUD*=1), statutory right of redemption is allowed (if *SRR*=1) and deficiency judgment is prohibited (if *NODJ*=1).
Owner-occupied	An indicator variable receiving value of 1 if the property is occupied by the owner and 0 otherwise.
Economic and market conditions	
Average local loan foreclosure (*FCrate*)	We define the FCrate dummies for the following ranges: *FCrate_d50* = $[0, 50^{th}\ percentile]$, *FCrate_d75* = $(50^{th}\ percentile, 75^{th}\ percentile]$, *FCrate_d90* = $(75^{th}\ percentile, 90^{th}\ percentile]$, *FCrate_d100* = $(90^{th}\ percentile, 100^{th}\ percentile]$.
Average house price gap (*GapHprice*)	Gap between local house price under a common market condition and under distress condition. We define the *GapHprice* dummies: *GapHprice_d50* = $[0, 50^{th}\ percentile]$, *GapHprice_d75* = $(50^{th}\ percentile, 75^{th}\ percentile]$, *GapHprice_d90* = $(75^{th}\ percentile, 90^{th}\ percentile]$, *GapHprice_d100* = $(90^{th}\ percentile, 100^{th}\ percentile]$.
Unemployment rate (*Unemployment rate*)	Quarterly unemployment rate (unemployment) is collected at country level from Bureau of Labor Statistics (BLS).
Real GDP growth rate (*Real growth rate*)	Real GDP growth rate (*Growth*) is calculated from quarterly real GDP collected at country level from Bureau of Economic Analysis (BEA).
Time to end of observation (*TimeToEOO*)	Time gap between default events and the time of last available loss information. This variable is exploited as a control for a possible loss information bias mentioned earlier.

Table 3: Descriptive statistics of explanatory variables

This table provides means and standard deviations (Std.Dev) of explanatory variables used in each equation. For a description of the variables, we refer to more details in Table 2.

Variable	Default equation		Cure equation		LGD equation	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
ARM	0.555	0.497	-	-	-	-
CLTV	0.742	0.255	0.913	0.262	0.943	0.255
Current IR	6.533	1.913	7.568	1.949	7.673	1.875
FCRate	0.016	0.012	0.026	0.016	0.028	0.016
FICO	677	73	643	67	646	66
GapHprice	0.790	0.347	0.926	0.352	0.982	0.339
Loan size	12.204	0.804	12.194	0.728	12.212	0.711
Modification	0.004	0.066	-	-	-	-
Owner Occupied	0.863	0.344	0.875	0.330	0.867	0.340
TimeToEOO	-	-	6.143	2.011	6.323	1.701
Unemployment rate	6.501	1.917	6.651	1.983	6.559	1.978
Real GDP growth rate	0.800	0.813	0.513	0.932	0.445	0.952
Loan age	3.877	3.100	3.142	2.262	2.827	1.889
ILR	0.566	0.496	0.648	0.478	0.625	0.484
NODJ	-	-	0.335	0.472	0.367	0.482
NOJUD	-	-	-	-	0.004	0.063
SRR	-	-	-	-	0.816	0.387
Observations	2,899,794		58,519		44,564	

3. Modelling framework

We develop a framework that captures the selection mechanism of the observed loss severities as follows: (i) the probability of default; (ii) the probability of cure for default events; and (iii) the non-zero loss severity for defaults and non-cures. This selection mechanism is visually shown in Figure 2.

To avoid the possible sample selection bias, which is due to a reduction in the sample from *N* loans to the reduced sample of defaulted loans, we control the observability by a bivariate Probit sample selection model of default and cure (see Boyes et al., 1989; Greene, 1998, for the bias issues, and Wolter and Rösch, 2014; Andersson and Mayock, 2014, for a similar treatment). Our general model framework is a system that includes three linear equations representing three selection steps mechanism described in the following.

3.1. Probability of default (PD)

$$D_{it}^{*} = X_{i,t-1}\beta + u_{it} \qquad u_{it} \sim N(0,1)$$

$$D_{it} = \begin{cases} 1 & if\ D_{it}^{*} > 0 \\ 0 & if\ D_{it}^{*} \leq 0 \end{cases}, \tag{4}$$

Here, we adopt a common approach of a PD estimation that is well developed in the literature. We assume D_{it}^{*} as the underlying latent process that defines whether loan *i* is defaulted at time *t*. If $D_{it}^{*} > 0$, a default occurs (i.e., $D_{it} = 1$), otherwise the loan is not defaulted (i.e., $D_{it} = 0$). $X_{i,t-1}$ collects all time-varying risk drivers observed in the previous quarter that explain the PD, β is the vector of parameters and the error term u_{it} is assumed to follow a standard normal distribution.

3.2 Probability of cure (PC)

$$C_{it}^{*} = \Theta_{i,t-1}\lambda + v_{it} \qquad v_{it} \sim N(0,1)$$

$$C_{it} = \begin{cases} 1 & if\ C_{it}^{*} > 0\ and\ D_{it} = 1 \\ 0 & if\ C_{it}^{*} \leq 0\ and\ D_{it} = 1 \end{cases}, \tag{5}$$

In our PC model (5), the cure variable, C_{it}, is only observable if the default occurs (i.e., $D_{it} = 1$). Similar to our PD model, C_{it} is a binary variable indicating whether the defaulted loan is cured ($C_{it} = 1$) or non-cured ($C_{it} = 0$), which is characterised by the underlying latent process $C_{it}^{*} > 0$ or $C_{it}^{*} \leq 0$, respectively. $\Theta_{i,t-1}$ is a vector of all time-varying variables observed in previous quarter that explains the PC, λ is the vector of parameters and the error term v_{it} is assumed to follow a standard normal distribution.

3.3 Loss Given Default

$$L_{it} = Z_{i,t-1}\alpha + \varepsilon_{it} \qquad \varepsilon_{it} \sim N(0, \sigma_{\varepsilon}^{2}) \tag{6}$$

Following the selection mechanism shown in Figure 2, the magnitude of non-zero loss severity regarding loan *i* at time *t*, L_{it}, is only observed if loan *i* defaults at time *t* (i.e., $D_{it} = 1$) and it is not cured (i.e., $C_{it} = 0$). In that case, the loss severity variable, L_{it}, follows a linear relationship shown in Eq. (6), where, $Z_{i,t-1}$ is a vector of all time-varying determinants of the LGD observed in the previous quarter and α is the associated parameter vector. We let the error term, ε_{it}, following a normal distribution with zero mean and σ_{ε}^{2} variance.

Since these three linear equations are included in one closed system, the vector of error terms follows a multivariate normal distribution as below

$$\begin{pmatrix} u_{it} \\ v_{it} \\ \varepsilon_{it} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{uv} & \rho_{u\varepsilon}\sigma_{\varepsilon} \\ \rho_{uv} & 1 & \rho_{v\varepsilon}\sigma_{\varepsilon} \\ \rho_{u\varepsilon}\sigma_{\varepsilon} & \rho_{v\varepsilon}\sigma_{\varepsilon} & \sigma_{\varepsilon}^2 \end{pmatrix} \right] \tag{7}$$

where ρ_{xy} denotes the correlation between variables *x* and *y*.

3.4 Model estimation

Following selection mechanism as in Figure 2, we can represent three combinations parametrically as follows[7],

a. If the loan *i* is not defaulted at time *t* ($D_{it} = 0$):

   $$\Pr(D_{it} = 0) = \Pr\left(D_{it}^* \leq 0 \middle| X_{i,t-1}\right) = 1 - \Phi\left(X_{i,t-1}\beta\right) \tag{8}$$

   where $\Phi(.)$ denotes the cumulative distribution function of the standardised normal distribution.

b. If the loan *i* is defaulted at time *t* but it is cured ($D_{it} = 1$ and $C_{it} = 1$):

   $$\Pr(D_{it} = 1, C_{it} = 1) = \Pr\left(D_{it}^* > 0, C_{it}^* > 0 \middle| X_{i,t-1}, \Theta_{i,t-1}\right) = \Phi_2\left(X_{i,t-1}\beta, \Theta_{i,t-1}\lambda, \rho_{uv}\right) \tag{9}$$

   where $\Phi_2(.)$ denotes the cumulative distribution function of standardised bivariate normal distribution.

c. If the loan *i* is defaulted at time *t* but it is non-cured ($D_{it} = 1$ and $C_{it} = 0$) then the L_{it} is observed and follows Eq. (6):

$$\Pr\left(L_{it}, D_{it} = 1, C_{it} = 0 \middle| X_{i,t-1}, \Theta_{i,t-1}, Z_{i,t-1}\right)$$
$$= \frac{1}{\sigma_{\varepsilon}} \phi\left(\frac{L_{it} - Z_{i,t-1}\alpha}{\sigma_{\varepsilon}}\right)$$
$$\cdot \Phi_2\left(\frac{X_{i,t-1}\beta + \frac{\rho_{u\varepsilon}}{\sigma_{\varepsilon}}\left(L_{it} - Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{u\varepsilon}^2}}, \frac{-\Theta_{i,t-1}\lambda - \frac{\rho_{v\varepsilon}}{\sigma_{\varepsilon}}\left(L_{it} - Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{v\varepsilon}^2}}, \rho_{uv|\varepsilon}^*\right) \tag{10}$$

where $\phi(.)$ denotes probability density function of standardised normal distribution, and,

$$\rho_{uv|\varepsilon}^* = \frac{-\rho_{uv} + \rho_{u\varepsilon}\rho_{v\varepsilon}}{\sqrt{1-\rho_{u\varepsilon}^2}\sqrt{1-\rho_{v\varepsilon}^2}}$$

This leads to the likelihood function of the model

[7] Proofs of missing links are provided in the Appendix A.

$$L = \prod_{t=1}^{T}\prod_{i=1}^{N}\left(1-\Phi\left(X_{i,t-1}\beta\right)\right)^{1-D_{it}} \cdot \left(\Phi_2\left(X_{i,t-1}\beta, \Theta_{i,t-1}\lambda, \rho_{uv}\right)\right)^{D_{it}\cdot C_{it}}$$

$$\cdot \left(\frac{1}{\sigma_\varepsilon}\phi\left(\frac{L_{it}-Z_{i,t-1}\alpha}{\sigma_\varepsilon}\right) \cdot \Phi_2\left(\frac{X_{i,t-1}\beta + \frac{\rho_{u\varepsilon}}{\sigma_\varepsilon}\left(L_{it}-Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{u\varepsilon}^2}}, \frac{-\Theta_{i,t-1}\lambda - \frac{\rho_{v\varepsilon}}{\sigma_\varepsilon}\left(L_{it}-Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{v\varepsilon}^2}}, \rho_{uv|\varepsilon}^*\right)\right)^{D_{it}\cdot(1-C_{it})} \quad (11)$$

We estimate the model by maximizing the log of the likelihood function (11) using the Dual Quasi-Newton optimisation technique.

3.5 Simulations

In this section we conduct some simulations to assess the finite sample performance of our model as well as the convergence of our estimators to their true values. We obtain the simulation results based on 100 replications for different choices of sample sizes (*T*), including 5000, 10,000, 20,000, 50,000 and 100,000 observations. The response variables, namely *D_i*, *C_i* and *L_i* are generated using the following data generating process:

$$D_i^* = a_0 + a_1X_{1i} + a_2X_{2i} + u_i$$

$$D_i = \begin{cases} 1 & if\ D_i^* > 0 \\ 0 & if\ D_i^* \leq 0 \end{cases},$$

$$C_i^* = b_0 + b_1X_{1i} + b_2X_{2i} + v_i$$

$$C_i = \begin{cases} 1 & if\ C_i^* > 0\ and\ D_i = 1 \\ 0 & if\ C_i^* \leq 0\ and\ D_i = 1 \end{cases},$$

$$L_i = c_0 + c_1X_{1i} + c_2X_{2i} + \varepsilon_i \text{ if } C_i = 0 \text{ and } D_i = 1,$$

$$\begin{pmatrix} u_i \\ v_i \\ \varepsilon_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{uv} & \rho_{u\varepsilon}\sigma_\varepsilon \\ \rho_{uv} & 1 & \rho_{v\varepsilon}\sigma_\varepsilon \\ \rho_{u\varepsilon}\sigma_\varepsilon & \rho_{v\varepsilon}\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix}\right],$$

where X_{1i} and X_{2i} are randomly generated from the standard normal distribution and the true parameters are set as, $a_0 = 0.5, a_1 = 0.2, a_2 = 0.6, b_0 = 0.2, b_1 = 0.5, b_2 = -0.3, c_0 = 0.4, c_1 = -0.1, c_2 = 0.7$. Vector of error terms $(u_i, v_i, \varepsilon_i)'$ are randomly drawn from the multivariate normal distribution with $\rho_{uv} = 0.5, \rho_{u\varepsilon} = 0.3, \rho_{v\varepsilon} = 0.6, \sigma_\varepsilon = 0.4$. Based on the simulated data of D_i, C_i and L_i as response variables and that of X_{1i} and X_{2i} as explanatory variable, we apply the estimation procedure outlined in subsection 3.2 and obtain the Mean Absolute Errors (MAE) and Root Mean Square Error (RMSE) of estimates of the parameters as shown in Table 4. As can be seen, our estimators work well in terms of both consistency and convergence properties, demonstrated by significant decreases in the deviation of estimates from the true values when the sample size increases. In this simulation exercise, we restrict the maximum sample size of 100,000 observations due to computer burden. It is, however, important to note that our empirical analyses presented in the following sections are based on millions of observations, making the estimation even more reliable.

Table 4: Mean absolute error (MAE) and Root mean square error (RMSE) of estimates from simulated data

This table reports the simulation results for a finite sample performance of our proposed model. The MAE and RMSE, calculated based on 100 replications for five different choices of sample sizes, indicate the deviation of our estimates from the true values of parameters. The results show significant decreases in the deviations when the sample size increases, an indication of well-performed estimators in terms of both consistency and convergence properties.

Parameters	True values	T=5,000		T=10,000		T=20,000		T=50,000		T=100,000	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
a_0	0.5	0.016	0.020	0.010	0.013	0.008	0.010	0.005	0.007	0.004	0.005
a_1	0.2	0.017	0.021	0.011	0.014	0.008	0.010	0.005	0.006	0.004	0.005
a_2	0.6	0.016	0.020	0.012	0.015	0.008	0.011	0.006	0.008	0.004	0.004
b_0	0.2	0.101	0.137	0.086	0.111	0.056	0.078	0.037	0.049	0.027	0.033
b_1	0.5	0.024	0.029	0.014	0.018	0.010	0.013	0.006	0.008	0.004	0.005
b_2	-0.3	0.067	0.084	0.054	0.068	0.035	0.048	0.024	0.031	0.017	0.021
c_0	0.4	0.100	0.141	0.074	0.110	0.048	0.086	0.025	0.032	0.019	0.023
c_1	-0.1	0.037	0.053	0.028	0.042	0.019	0.033	0.007	0.009	0.006	0.007
c_2	0.7	0.035	0.046	0.026	0.034	0.020	0.027	0.013	0.016	0.009	0.011
$\rho_{u,v}$	0.5	0.161	0.223	0.127	0.170	0.088	0.121	0.056	0.073	0.039	0.049
$\rho_{u,\varepsilon}$	0.3	0.265	0.357	0.191	0.259	0.142	0.210	0.087	0.114	0.061	0.083
$\rho_{v,\varepsilon}$	0.6	0.223	0.367	0.179	0.298	0.113	0.238	0.034	0.044	0.028	0.037
σ_ε	0.4	0.024	0.031	0.021	0.026	0.014	0.018	0.008	0.010	0.006	0.008

4. Empirical results

We base our model estimation on a 10% random sample of the main dataset to ease computation.[8] We omitted the missing observations in main variables and risk drivers. For empirical analysis and robustness check, we employ three alternative specifications for our selection model and its comparative models, named as the Restricted, Unrestricted and Non-linear specifications. These three specifications differ in terms of different sets of explanatory variables used for modelling (see Table 7, Table 8 and Table 9 for a detail set of variables). While the Restricted version only includes main drivers, the Unrestricted version expands by controlling macroeconomic variables (real GDP growth rate and unemployment rate) and the vintage effects. In addition, the non-linear version provides a robustness check of the existence of non-linear regional foreclosure contagion effect by adding to the Unrestricted version the dummies of house price under distress and regional foreclosure rate.

4.1 Model performance

We first verify the necessity of the dependent structure in our model by comparing it with the independent structure that prohibits the error correlation structures among PD, PC and non-zero LGDs. We define the two models as follows,

- **Dependent Model:** the three stage selection regression model derived in this paper (see Eq. (11)) assuming dependence between these three stochastic processes;
- **Independent Model:** separate regression models for the PD, PC and non-zero LGDs but assuming independence between these three stochastic processes. This model is equivalent with the independent estimation of a probit model for default, a probit model for cure events and an OLS model for non-zero LGDs.

We perform the likelihood ratio test with the null hypothesis of the independent model against the alternative hypothesis of the dependent model. As shown in Table 5, we observe a strong rejection of the independent structure, suggesting a necessity of a dependent structure in our selection model for the data sample. This is further supported by statistically significant pairwise error correlation between cure and non-zero LGD equations[9]. However, we note that the magnitude of the pairwise error correlations is small, which is due to a large number of statistically significant drivers included in the models. As can be seen, the magnitude of the

[8] We have confirmed robustness using a bootstrap. Results are available on request.

[9] Note that the pairwise correlation structures among default rate, cure rate and non-zero LGD *in their mean level* are not necessarily equivalent with their pairwise *error* correlation structures since parts of their variation are captured by the explanatory variables.

pairwise error correlations becomes less significant when the number of explanatory variables increase. However, it is worth noting that models with a higher number of explanatory variables do not necessarily provide a better predictive quality than one with a lower number of explanatory variables, to that extent we will discuss later in the out-of-time predictive performance part. Hence, our model with the dependent structure remains its usefulness, particularly in the case when a limited number of explanatory variables are found.

Table 5: Estimates of correlation and variance parameters

This table reports the correlation and variance estimates of three alternative models that we employed for empirical analysis and robustness check. Details of explanatory variables of *Restricted*, *Unrestricted* and *Non-linear* models can be found in Table 7, 8 and 9. Standard errors are reported in parentheses. ***, ** and * denote the estimates are statistically significant at 1%, 5% and 10% level of significance, respectively.

Parameter	**Restricted**		**Unrestricted**		**Non-linear**	
	Dependent	**Independent**	**Dependent**	**Independent**	**Dependent**	**Independent**
Error correlation (default, cure), $\rho_{u,v}$	0.059		0.062		0.053	
	(0.075)		(0.089)		(0.074)	
Error correlation (default, LGD), $\rho_{u,\varepsilon}$	0.062		0.022		0.041	
	(0.045)		(0.048)		(0.047)	
Error correlation (cure, LGD), $\rho_{v,\varepsilon}$	0.087***		0.065**		-0.004	
	(0.029)		(0.029)		(0.027)	
σ_{ε}	0.235***	0.234***	0.234***	0.233***	0.235***	0.234***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
-2 Log Likelihood	558992	559014	558309	558335	558989	558993
LR test Independent vs. Dependent	$\chi^2_{statistic}$ = 22		$\chi^2_{statistic}$ = 26		$\chi^2_{statistic}$ = 4	
	$p\text{-}value$=0.0000		$p\text{-}value$=0.0000		$p\text{-}value$=0.2615	

We assess the performance of our model (both dependent and independent structure) with the OLS model (a single OLS regression model for both zero and non-zero LGDs), which is the current practice in the literature and industry. Regarding the in-sample goodness of fit statistics, we observe a comparable performance of the dependent and independent structure across all equations examined (see Table 7, Table 8 and Table 9). Meanwhile, the OLS model provides a slightly better fit for the LGDs than the dependent and independent structure in terms of Adjusted R-square. This is, however, not surprising, as the OLS model uses both zero and non-zero LGDs observations, whereas, the dependent and independent structure of our model only consider the non-zero LGDs and control for selection biases.

We assess the out-of-time predictive accuracy of our proposed model with the benchmark OLS model. We find that the Restricted specification of the OLS model provides the best predictive performance in terms of Root Mean Square Errors (RMSE, hereafter) compared to its Unrestricted and Non-linear specification[10]. Therefore, we employ the Restricted specification for the out-of-time predictive accuracy comparison between our proposed model and the benchmark OLS model. We evaluate the forecast accuracy in five consecutive years from 2008 to 2012[11]. We report the RMSE statistic as a mean for measuring the out-of-time forecasting accuracy of considered models as well as the difference in predictive performance of our proposed models with the benchmark OLS model in Table 6. For a statistical inference, we perform the two-sample *t*-test for difference in mean of the square errors of the predictions between Dependent model and Independent/OLS models.

[10] We do not report the RMSEs of the Unrestricted and Non-linear specification of the OLS model to conserve space. However, details are available upon request.

[11] The reason for not considering years from 2013 to 2015 is that most of the non-zero losses arose within three years of the default events (see Figure 4). In empirical modeling, we employ the *TimeToEOO* variable to control an effect of time gap between default events and the time that the last loss was observed. However, in prediction, the information on *TimeToEOO* is not available, we, therefore, drop this variable and use 2012 as the final year for forecasting evaluation to avoid the potential loss information bias.

Table 6: Out-of-time predictive performance statistics

This table shows the Root Mean Square Errors (RMSE) of the Out-of-time prediction for 3 selected models as well as the relative difference between RMSEs of the Dependent model's out-of-time prediction and that of other comparative models. *OLS*, the one-step LGD OLS regression model, is chosen as the benchmark model for a comparison purpose due to its popularity in the literature and industry. In this assessment, we chose the *restricted* form for all three models since we found that the *restricted* form of the benchmark OLS outperforms its *unrestricted* and *non-linear* form in terms of out-of-time prediction. The OLS model has the specification as, $L_{it}^{*} = Z_{i,t-1}\alpha^{*} + \varepsilon_{it}^{*}$, in which L_{it}^{*} includes both observations of zero and non-zero LGDs. The table also reports the two-sample *t*-test for difference in mean of the square errors of the predictions between Dependent model and Independent/OLS model. The null hypothesis, H_0*: Mean Square Error of Dependent predictions = Mean Square Error of Independent/OLS predictions*, is tested against, H_A*: Mean Square Error of Dependent predictions < Mean Square Error of Independent/OLS predictions*. T-statistics are reported in parentheses. ***, ** and * indicate rejections of the null hypothesis H_0 at 1%, 5% and 10% level of significance, respectively.

	RMSEs of Out-of-Time prediction			**ΔRMSE between Dependent model and others**	
	Dependent	**Independent**	**OLS**	**Independent**	**OLS**
Out of sample 2008					
In-sample 2004-2007	0.2737	0.2768	0.2862	-1.11%	-4.36%
				(-1.581)*	(-5.586)***
Out of sample 2009					
In-sample 2004-2008	0.2844	0.2856	0.2867	-0.41%	-0.82%
				(-0.664)	(-1.151)
Out of sample 2010					
In-sample 2004-2009	0.3103	0.3169	0.3130	-2.09%	-0.86%
				(-0.286)	(-0.165)
Out of sample 2011					
In-sample 2004-2010	0.2900	0.2912	0.2904	-0.42%	-0.13%
				(-0.436)	(-0.171)
Out of sample 2012					
In-sample 2004-2011	0.2796	0.2813	0.2802	-0.60%	-0.19%
				(-0.351)	(-0.186)

It is clear that our proposed dependent framework performs better than the Independent model and the current benchmark model in the literature and industry (i.e., the OLS model) as evidenced by its consistent lower RMSEs in all cases examined. More importantly, our model shows nearly (and statistically significant) 4.5% better predictive quality than the OLS model during the GFC. This is mainly due to the outperformance of our proposed model in forecasting the high-value ranges of LGDs in a comparison with the OLS model in 2008 (see Figure 5). Part of the LGD variation is omitted when the uncertainty of the default and cure events is not considered in the OLS model. This problem may be more significant during the turbulence periods when variables tend to be highly correlated so that we can observe a remarkable difference in forecasting performance as mentioned above. We support this argument by respectively plotting the average actual LGDs versus out-of-time predicted LGDs according fifteen equal groups[12] of the CLTV variable – the most significant determinant of PD in our model (see Figure 6), and according to fifteen equal groups of the *GapHprice* variable – the most significant determinant of PC in our model[13] (see Figure 7).

[12] The choice of fifteen groups is subjective in the essence to balance the visualisation and information delivered through the graph.

[13] We determine the most significant explanatory variables by looking at their associated highest test statistics in the estimated outputs of the models.

Figure 5: Average actual LGDs versus out-of-time predicted LGDs by groups of predicted LGD

This figure plots average actual LGDs versus out-of-time predicted LGDs generated from three selected models for fifteen equal groups of predicted LGDs. *Model 1* represents the Dependent Model, *Model 2* denotes the Independent Model and *Model 3* indicates OLS Model. The diagonal line indicates a perfect prediction, which means the predicted values match actual values exactly. Note that there is no requirement in an out-of-sample analysis that the scatters are centred around the diagonal.

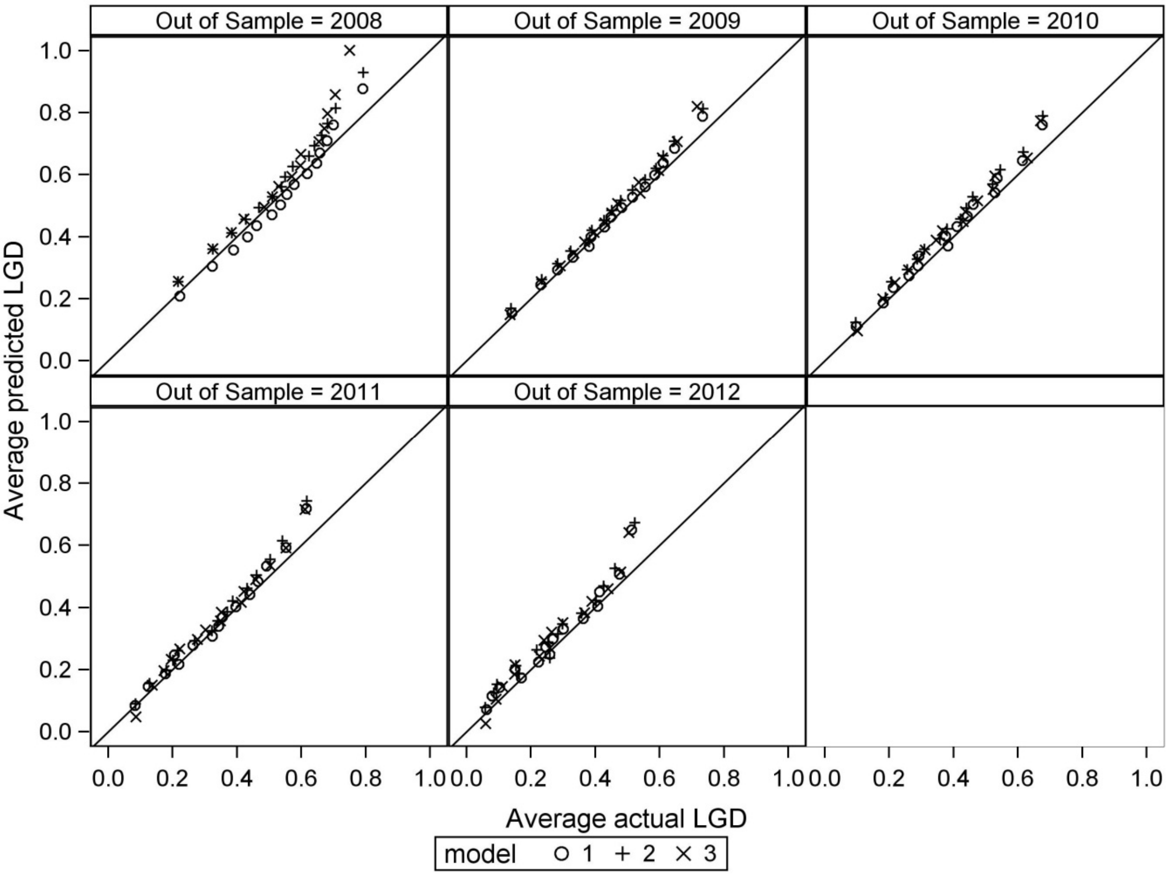


Figure 6: Average actual LGDs versus out-of-time predicted LGDs by groups of CLTV

This figure plots average actual LGDs versus out-of-time predicted LGDs generated from three selected models for fifteen equal groups of CLTV (the most significant explanatory variable for PD). Model 1 represents the Dependent Model, Model 2 denotes the Independent Model and Model 3 indicates OLS Model. The diagonal line indicates a perfect prediction, which means the predicted values match actual values exactly. Note that there is no requirement in an out-of-sample analysis that the scatters are centred around the diagonal.

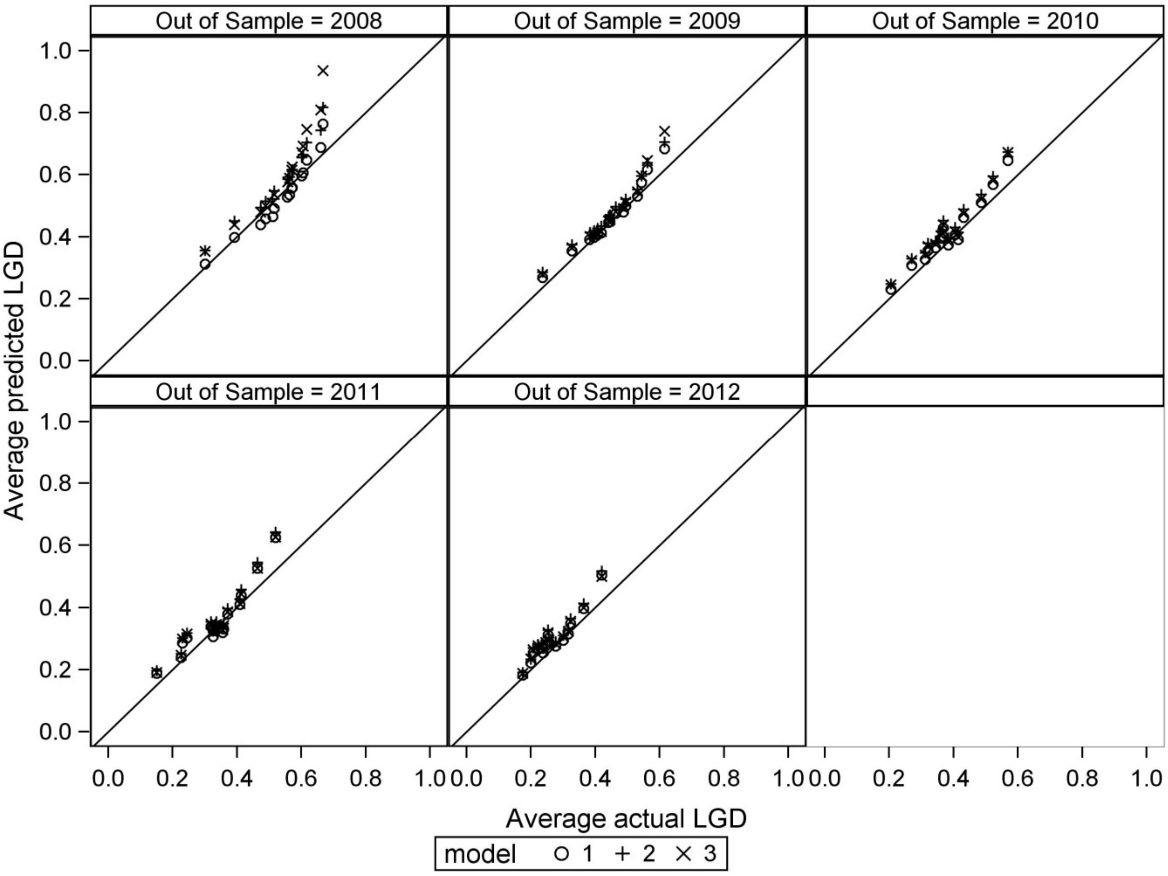
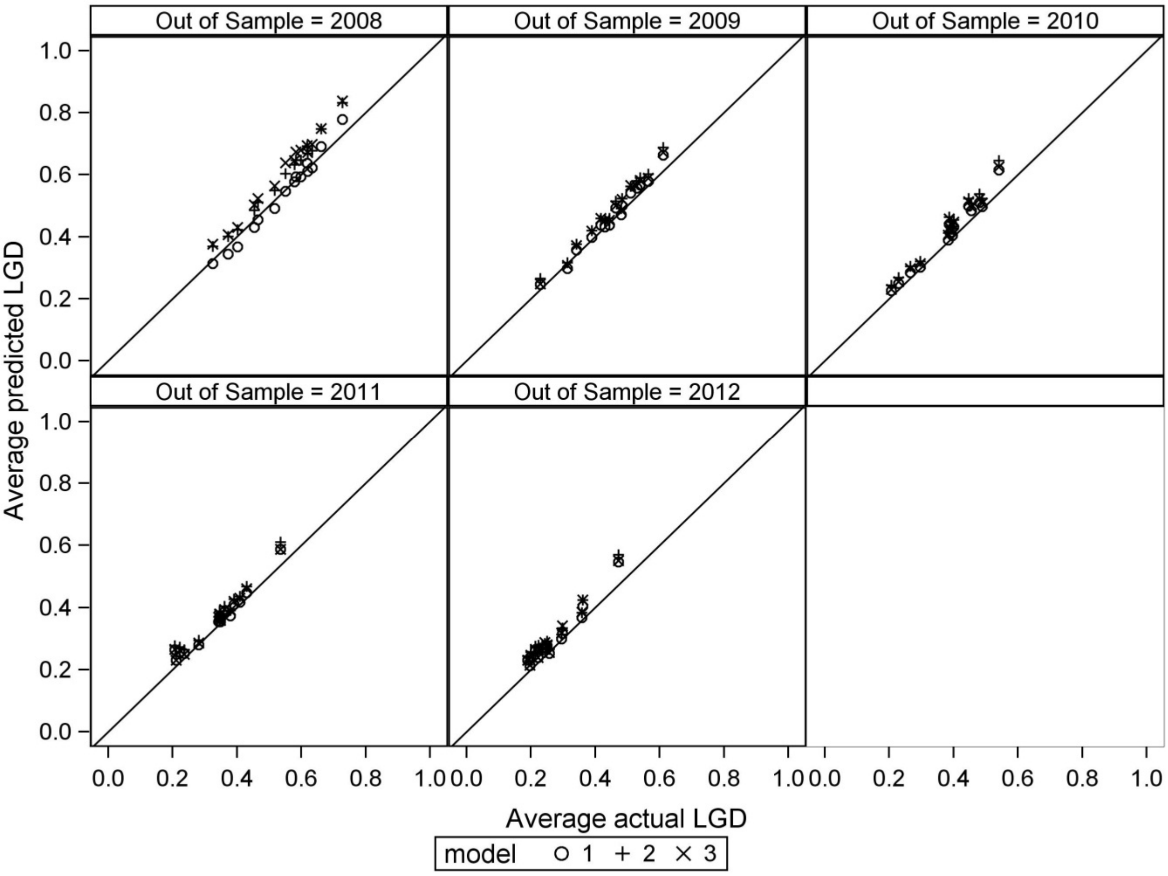


Figure 7: Average actual LGDs versus out-of-time predicted LGDs by groups of GapHprice

This figure plots average actual LGDs versus out-of-time predicted LGDs generated from three selected models for fifteen equal groups of GapHprice (the most significant explanatory variable for PC). Model 1 represents the Dependent Model, Model 2 denotes the Independent Model and Model 3 indicates OLS Model. The diagonal line indicates a perfect prediction, which means the predicted values match actual values exactly. Note that there is no requirement in an out-of-sample analysis that the scatters are centred around the diagonal.



It can be seen that our proposed dependent model provides a superior forecast performance to that of OLS regression in the high-value ranges of LGDs, which is associated with the high-value group of CLTV (equivalently, high-value groups of PD) and *GapHprice* (equivalently, low-value groups of PC). Notably, this difference in the models' predictive performance is only remarkable in the GFC but it becomes less and less significant in the following years. This fact is consistent with our claim that an application of the dependent structure among PD, PC and non-zero LGDs is essential to produce more precise forecasts, especially under the financial turbulence, when those credit risk measures tend to be highly correlated. Overlooking this structure may lead to a larger bias in prediction, especially for predicting the downturn LGD.

4.2 Probability of Default (PD) equation

We present the estimated results for the PD equation in Table 7. Our results are robust across the models examined and consistent with the economic intuitions and empirical findings in the literature. We find that PD is positively associated with CLTV and current interest rate, which is strongly supported by previous studies (see Amromin and Paulson, 2009; Elul et al., 2010; Goodman et al., 2010, for evidence on CLTV and Demyanyk and Hemert, 2011, for evidence on current interest rate). Consistent with Amromin and Paulson (2009), our result indicates that adjustable-rate mortgage loans are riskier, and hence, expose a higher likelihood of default than the fixed-rate mortgage loans. As expected, the residential mortgage loans for owner-occupied purpose are less risky than other purpose of occupancy (such as investment). It is not surprising that mortgage loans with a higher FICO score are less likely to default since a higher FICO score represents a better credit quality. This result has been widely documented in the literature (see Elul et al., 2010, and Demyanyk and Hemert, 2011, among others). Further, consistent with previous studies (e.g., Elul et al., 2010; Demyanyk, Koijen and Van Hemert, 2011; Gerardi et al., 2013; and Campbell and Cocco, 2015), loan associated with insufficient repayment in the previous quarter are more likely to default. Other factors including loan size and loan age show non-linear effects on mortgage default, which are also in line with the expectation (see Li et al., 2012, for an example of loan size and Bajari et al., 2008, for an example of loan age). We find that modified loans are less likely to be foreclosed. This may be due to the common bank's strategies in working out the delinquent loans. For example, at early stages of delinquency, banks normally support borrowers to correct their loans by providing various types of modifications such as allowing the customer to miss a few payments (while capitalising interest). This would help borrowers to overcome short-term liquidity constraints and keep their loans away from the foreclosure process.

Table 7: Estimates of probability of default equation

This table reports the estimation outputs of the probability of default equation for Dependent and Independent models under their three forms, *Restricted*, *Unrestricted* and *Non-linear*, which are differed in terms of explanatory variables. For a description of the variables, we refer to more details in Table 2. FICO is divided by 1000, FCrate is multiplied by 10, Loan age, Loan size Current IR, Unemployment rate and Real growth rate are divided by 10. Standard errors are reported in parentheses. ***, ** and * denote the estimates are statistically significant at 1%, 5% and 10% level of significance, respectively. AUROC denotes the Area Under the Receiver Operating Characteristics curve, which is a popular measure for discrimination in credit risk.

Parameter	Restricted		Unrestricted		Non-linear	
	Dependent	Independent	Dependent	Independent	Dependent	Independent
FICO	-1.921***	-1.925***	-1.942***	-1.945***	-1.964***	-1.966***
	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
CLTV	0.808***	0.808***	0.756***	0.757***	0.813***	0.813***
	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
GapHprice	0.084***	0.084***	0.091***	0.091***	-	-
	(0.007)	(0.007)	(0.007)	(0.007)	-	-
GapHprice_d75	-	-	-	-	0.04***	0.04***
	-	-	-	-	(0.005)	(0.005)
GapHprice_d90	-	-	-	-	0.063***	0.063***
	-	-	-	-	(0.006)	(0.006)
GapHprice_d100	-	-	-	-	0.068***	0.068***
	-	-	-	-	(0.007)	(0.007)
Fcrate	1.224***	1.224***	1.139***	1.14***	-	-
	(0.015)	(0.015)	(0.017)	(0.017)	-	-
Fcrate_d75	-	-	-	-	0.178***	0.178***
	-	-	-	-	(0.005)	(0.005)
Fcrate_d90	-	-	-	-	0.291***	0.291***
	-	-	-	-	(0.006)	(0.006)
Fcrate_d100	-	-	-	-	0.456***	0.456***
	-	-	-	-	(0.007)	(0.007)
Current IR	0.906***	0.905***	0.898***	0.897***	0.886***	0.886***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
ARM	0.225***	0.225***	0.23***	0.23***	0.226***	0.226***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Owner-occupied	-0.093***	-0.093***	-0.09***	-0.09***	-0.092***	-0.092***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Modification	-0.29***	-0.296***	-0.314***	-0.318***	-0.309***	-0.313***

	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
ILR	0.165***	0.164***	0.16***	0.16***	0.156***	0.156***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Loan age	0.132***	0.13***	0.084***	0.085***	0.054**	0.055**
	(0.019)	(0.019)	(0.022)	(0.022)	(0.022)	(0.022)
Loan age square	-0.072***	-0.072***	-0.036**	-0.037**	-0.011	-0.012
	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)	(0.016)
Loan size	5.147***	5.148***	5.114***	5.114***	5.744***	5.745***
	(0.645)	(0.626)	(0.646)	(0.628)	(0.649)	(0.632)
Loan size square	-1.723***	-1.723***	-1.737***	-1.736***	-2.005***	-2.005***
	(0.263)	(0.256)	(0.264)	(0.257)	(0.265)	(0.258)
Real growth rate	-	-	-0.261***	-0.261***	-0.319***	-0.319***
	-	-	(0.023)	(0.023)	(0.023)	(0.023)
Unemployment rate	-	-	0.076***	0.076***	0.048***	0.048***
	-	-	(0.014)	(0.014)	(0.014)	(0.014)
Origin years	-	-	Yes	Yes	Yes	Yes
AUROC	0.788	0.788	0.788	0.788	0.787	0.787
Pseudo R-Square	11.4%	11.4%	11.5%	11.5%	11.5%	11.5%
Observations	2,899,794	2,899,794	2,899,794	2,899,794	2,899,794	2,899,794

Particularly, we find that an increase in the average local foreclosure rate in the previous quarter significantly increases the current likelihood of default of residential mortgage loans. Our result may be explained by theories of observational learning (e.g., Agarwal et al., 2012) and behavioural responses (e.g., Seiler et al., 2014). An observation of high local foreclosure rate may lead borrowers to further devalue their property and strengthen their belief on a declining housing market. Furthermore, this observation may also facilitate the dishonouring behaviour of borrowers towards default decisions. This belief/behaviour may finally trigger a higher local default rate. A similar intuition also applies to the positive effect of average gap between local house price under common market condition and under distress condition on the PD. The higher devaluation of local house price under distress compared to common market condition (due to poor maintenance and various types of discounts during the repossession process) may lead to borrowers feeling more certain about the decreasing housing market. This may in turn lessen borrowers' incentive to keep the loans.

4.3 Probability of Cure (PC) equation

Estimated results of the PC equation presented in Table 8 have shown some interesting features. Firstly, the foreclosed loans with higher FICO scores are less likely to be cured. FICO (2008) research shows that higher FICO scores tend to be more stable over time, which means higher FICO customers have put more effort into fulfilling their payment obligations. When high FICO customers face constraints (e.g., divorce, short-term unemployment or demotion) and miss a loan/interest repayment, they would try their best to rectify the loans at an early stage of delinquency. They would only give up and allow the loans to be foreclosed if they understood that it is not worthwhile maintaining the obligations and their FICO profiles. This finally leads to the losses being less likely to be fully recovered.

Table 8: Estimates of probability of cure equation

This table reports the estimation outputs of the probability of cure equation for Dependent and Independent models under their three forms, *Restricted*, *Unrestricted* and *Non-linear*, which are differed in terms of explanatory variables. For a description of the variables, we refer to more details in Table 2. FICO is divided by 1000, FCrate is multiplied by 10, Loan age, Loan size Current IR, Unemployment rate and Real growth rate are divided by 10. Standard errors are reported in parentheses. ***, ** and * denote the estimates are statistically significant at 1%, 5% and 10% level of significance, respectively. AUROC denotes the Area Under the Receiver Operating Characteristics curve, which is a popular measure for discrimination in credit risk.

Parameter	**Restricted**		**Unrestricted**		**Non-linear**	
	Dependent	**Independent**	**Dependent**	**Independent**	**Dependent**	**Independent**
FICO	-1.387***	-1.36***	-1.345***	-1.32***	-1.27***	-1.252***
	(0.168)	(0.111)	(0.19)	(0.112)	(0.171)	(0.112)
CLTV	-1.194***	-1.207***	-1.218***	-1.231***	-1.148***	-1.153***
	(0.069)	(0.034)	(0.077)	(0.034)	(0.068)	(0.033)
GapHprice	-1.085***	-1.088***	-1.07***	-1.074***	-	-
	(0.028)	(0.026)	(0.029)	(0.026)	-	-
GapHprice_d75	-	-	-	-	-0.38***	-0.38***
	-	-	-	-	(0.017)	(0.017)
GapHprice_d90	-	-	-	-	-0.573***	-0.576***
	-	-	-	-	(0.021)	(0.02)
GapHprice_d100	-	-	-	-	-0.753***	-0.757***
	-	-	-	-	(0.026)	(0.025)
Fcrate	-0.135	-0.155***	0.071	0.051	-	-
	(0.096)	(0.058)	(0.108)	(0.065)	-	-
Fcrate_d75	-	-	-	-	-0.202***	-0.201***
	-	-	-	-	(0.022)	(0.018)
Fcrate_d90	-	-	-	-	-0.243***	-0.25***
	-	-	-	-	(0.029)	(0.02)
Fcrate_d100	-	-	-	-	-0.19***	-0.21***
	-	-	-	-	(0.039)	(0.025)
Current IR	-0.36***	-0.362***	-0.347***	-0.349***	-0.344***	-0.341***
	(0.073)	(0.039)	(0.084)	(0.04)	(0.072)	(0.04)
TimeToEOO	-0.105**	-0.089*	-0.443***	-0.429***	-0.439***	-0.428***
	(0.053)	(0.053)	(0.106)	(0.106)	(0.106)	(0.106)
Owner-Occupied	0.11***	0.114***	0.112***	0.116***	0.109***	0.117***
	(0.021)	(0.021)	(0.022)	(0.021)	(0.021)	(0.021)
ILR	0.039**	0.034**	0.036**	0.029**	0.04**	0.038***

	(0.017)	(0.014)	(0.018)	(0.014)	(0.016)	(0.014)
NODJ	-0.314***	-0.325***	-0.317***	-0.327***	-0.291***	-0.297***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Loan age	0.678***	0.702***	0.368**	0.388***	0.552***	0.567***
	(0.08)	(0.08)	(0.143)	(0.143)	(0.143)	(0.143)
Loan age square	-0.195***	-0.177***	-0.144**	-0.129**	-0.222***	-0.208***
	(0.05)	(0.051)	(0.061)	(0.061)	(0.061)	(0.061)
Loan size	-4.886**	-4.854***	-3.957**	-3.927**	-4.882**	-4.859***
	(1.989)	(1.844)	(1.951)	(1.831)	(1.993)	(1.872)
Loan size square	2.392***	2.422***	2.03**	2.06***	2.463***	2.486***
	(0.812)	(0.757)	(0.796)	(0.752)	(0.813)	(0.769)
Real growth rate	-	-	0.655***	0.665***	0.477***	0.482***
	-	-	(0.083)	(0.08)	(0.083)	(0.08)
Unemployment rates	-	-	-0.034	-0.018	0.052	0.064
	-	-	(0.046)	(0.045)	(0.047)	(0.047)
Origin years	-	-	Yes	Yes	Yes	Yes
AUROC	0.764	0.764	0.765	0.765	0.764	0.764
Pseudo R-Square	15.1%	15.1%	15.3%	15.3%	15.1%	15.1%
Observations	58,519	58,519	58,519	58,519	58,519	58,519

We do not find robust evidence about an effect of average local foreclosure rate on the probability of cure. However, it is interesting to normally observe a negative sign of an effect of average local foreclosure rate on the probability of cure (see Restricted and Non-linear specifications in Table 8). This result is consistent with our expectation since the local foreclosure rate is documented to exacerbate the local house price, and hence increase the probability of a collateral insufficiency (e.g., Campbell et al., 2011; Anenberg and Kung, 2014; Gerardi et al., 2015; and Guren and McQuade, 2015). Using a similar informational transmission intermediate as in the case of local foreclosure rate, an increase in the ratio of house price under common market condition to house price under distress condition in previous quarter statistically and significantly decreases the probability of cure of a foreclosed loan. In addition, we find that the foreclosed loan is more likely to be cured if it is associated with an event of insufficient loan repayment in the previous quarter. This is consistent with our expectation since insufficient loan repayment does not necessarily reflect a shortfall in the collateral value. This conjecture is supported by an insignificant effect of insufficient loan repayment event on the non-zero LGDs (see Table 9).

It is worth noting that our findings in terms of the nature of relationship among variables (except for the effect of average local foreclosure rate on the probability of cure) are robust across models examined.

4.4 Non-zero loss severity equation

We present the estimated results of non-zero LGD equation in Table 9. As suggested by the dependent and independent structure of our proposed approach, we do not find a significant association between loss severity and credit quality (FICO) scores. This is consistent with our expectation since it is a common agreement in credit risk area that FICO scores are default predictors yet not informative for predictions of loss severity. The OLS model produces an extraordinary result of a positive relationship between FICO and LGD (i.e., defaulted loans with higher credit quality generate higher loss severity). This might be due to the fact that the OLS model has used both cured and non-cured loans for estimation, and hence caused a bias result of FICO-LGD relationship since the FICO has an effect on the probability of cure (as discussed earlier) but not the non-zero LGD.

Effects of loan characteristics on non-zero LGD are also found to be consistent with existing literature and economic intuition. We find a strong positive relationship between CLTV and the non-zero LGD, which is strongly supported by previous empirical findings (e.g., Pennington-Cross, 2003; Calem and LaCour-Little, 2004; and Qi and Yang, 2009). This is also in line with the economic interpretation that a higher CLTV loan is associated with lower home equity value, which leads to lower recovery rate (i.e., 1-LGD) and, therefore, higher non-zero LGD (see Qi and Yang, 2009). As expected, we find loans with a higher interest rate are associated with higher non-zero LGD, which supports Zhang et al. (2010); whereas, defaulted loans with a higher age are associated with lower non-zero LGD (e.g., Lekkas et al., 1993; Pennington-Cross, 2003; Qi and Yang, 2009). Further, the relationship between non-zero LGD and loan size exhibits a convex parabola, which is similar to findings of Pennington-Cross (2003) and Calem and LaCour-Little (2004). This finding can be explained by the fixed costs that arise during the foreclosure process such as attorney and title fees, which may lead to a lower loss rate for a loan with a larger size. However, when a loan reaches a certain amount, it may face a higher risk for additional loan increment such as risk of greater house value depreciation[14].

[14] We take a further analysis (two sample *t*-test for difference in population mean) and find that the average house value depreciation of loan with size greater than the turning points of the convex parabola relationship (subjects to the models and ranges between 13 and 15) is statistically significantly larger than that of loan with size smaller than the turning points at 1% significant level. Detail is available upon request.

Table 9: Estimates of non-zero loss severity equation

This table reports the estimation outputs of the non-zero loss severity equation for Dependent, Independent and OLS models under their three forms, *Restricted*, *Unrestricted* and *Non-linear*, which are differed in terms of explanatory variables. For a description of the variables, we refer to more details in Table 2. FICO is divided by 1000, FCrate is multiplied by 10, Loan age, Loan size Current IR, Unemployment rate and Real growth rate are divided by 10. Standard errors are reported in parentheses. ***, ** and * denote the estimates are statistically significant at 1%, 5% and 10% level of significance, respectively. AUROC denotes the Area Under the Receiver Operating Characteristics curve, which is a popular measure for discrimination in credit risk.

Parameter	Restricted			Unrestricted			Non-linear		
	Dependent	Independent	OLS	Dependent	Independent	OLS	Dependent	Independent	OLS
FICO	-0.017	0.027	0.22***	0.002	0.026	0.213***	0.017	0.025	0.202***
	(0.028)	(0.021)	(0.021)	(0.029)	(0.021)	(0.022)	(0.029)	(0.021)	(0.022)
CLTV	0.314***	0.314***	0.379***	0.289***	0.302***	0.371***	0.299***	0.301***	0.359***
	(0.01)	(0.006)	(0.006)	(0.01)	(0.006)	(0.006)	(0.01)	(0.006)	(0.006)
GapHprice	0.264***	0.27***	0.348***	0.27***	0.273***	0.351***	-	-	-
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	-	-	-
GapHprice_d75	-	-	-	-	-	-	0.085***	0.087***	0.11***
	-	-	-	-	-	-	(0.004)	(0.003)	(0.003)
GapHprice_d90	-	-	-	-	-	-	0.151***	0.153***	0.192***
	-	-	-	-	-	-	(0.004)	(0.004)	(0.004)
GapHprice_d100	-	-	-	-	-	-	0.217***	0.22***	0.278***
	-	-	-	-	-	-	(0.005)	(0.004)	(0.005)
Fcrate	0.059***	0.065***	0.065***	0.053***	0.066***	0.042***	-	-	-
	(0.014)	(0.01)	(0.01)	(0.015)	(0.01)	(0.011)	-	-	-
Fcrate_d75	-	-	-	-	-	-	0.024***	0.02***	0.039***
	-	-	-	-	-	-	(0.004)	(0.004)	(0.004)
Fcrate_d90	-	-	-	-	-	-	0.059***	0.055***	0.077***
	-	-	-	-	-	-	(0.005)	(0.004)	(0.004)
Fcrate_d100	-	-	-	-	-	-	0.066***	0.064***	0.078***
	-	-	-	-	-	-	(0.006)	(0.005)	(0.005)
NOJUD	-0.019	-0.02	-0.028*	-0.022	-0.022	-0.03*	-0.024	-0.027	-0.035**
	(0.018)	(0.017)	(0.017)	(0.018)	(0.017)	(0.017)	(0.018)	(0.018)	(0.017)
SRR	0.049***	0.048***	0.029***	0.049***	0.048***	0.031***	0.047***	0.045***	0.026***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
NODJ	0.006**	0.005*	0.046***	0.008***	0.006**	0.047***	-0.005	-0.007**	0.034***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Current IR	0.182***	0.162***	0.17***	0.174***	0.16***	0.167***	0.191***	0.163***	0.167***

	(0.012)	(0.008)	(0.008)	(0.012)	(0.008)	(0.008)	(0.012)	(0.008)	(0.008)
TimeToEOO	-0.25***	-0.277***	-0.16***	-0.039	-0.115***	-0.03	-0.015	-0.085***	0.001
	(0.012)	(0.012)	(0.011)	(0.025)	(0.025)	(0.022)	(0.025)	(0.025)	(0.022)
Owner-occupied	-0.06***	-0.053***	-0.066***	-0.059***	-0.053***	-0.066***	-0.06***	-0.053***	-0.066***
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
ILR	-0.004	-0.003	-0.002	-0.003	-0.002	-0.001	-0.004	-0.003	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Loan age	-0.273***	-0.309***	-0.292***	-0.026	-0.147***	-0.171***	-0.076**	-0.172***	-0.21***
	(0.017)	(0.017)	(0.016)	(0.033)	(0.033)	(0.029)	(0.033)	(0.033)	(0.029)
Loan age square	0.133***	0.133***	0.105***	0.055***	0.095***	0.083***	0.086***	0.111***	0.102***
	(0.013)	(0.013)	(0.01)	(0.016)	(0.016)	(0.012)	(0.016)	(0.016)	(0.012)
Loan size	-5.219***	-12.934***	-5.36***	-5.159***	-12.637***	-5.281***	-4.964***	-12.401***	-5.055***
	(0.45)	(0.438)	(0.371)	(0.453)	(0.442)	(0.373)	(0.454)	(0.443)	(0.374)
Loan size square	1.723***	4.872***	1.843***	1.702***	4.751***	1.809***	1.616***	4.648***	1.703***
	(0.184)	(0.18)	(0.152)	(0.185)	(0.181)	(0.153)	(0.185)	(0.181)	(0.154)
Real growth rate	-	-	-	0.044***	0.037***	-0.047***	0.079***	0.073***	0.005
	-	-	-	(0.014)	(0.013)	(0.014)	(0.014)	(0.013)	(0.015)
Unemployment rate	-	-	-	0.045***	0.04***	0.037***	0.048***	0.047***	0.04***
	-	-	-	(0.01)	(0.009)	(0.009)	(0.01)	(0.01)	(0.009)
Origin Years	-	-	-	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R-square	29.4%	29.8%	30.9%	29.6%	30.0%	31.1%	29.1%	29.6%	30.6%
Observations	44,564	44,564	58,519	44,564	44,564	58,519	44,564	44,564	58,519

Regarding state foreclosure laws on property, we find that states with statutory right of redemption have higher non-zero LGD by 4.7% than other states. This result is consistent with previous studies (e.g., Clauretie and Herzog, 1990; Crawford and Rosenblatt, 1995; and Qi and Yang, 2009), and it can be explained that statutory right of redemption prolongs the foreclosure and liquidation process, which may finally lead to larger loss severity. In terms of other property characteristics, it is not surprising that owner-occupied properties experience a lower loss severity of between 5-6% compared to other owner occupancy purposes.

Furthermore, we find evidence that higher local foreclosure rate increases the non-zero LGD significantly (for every 1% increase in the average local foreclosure rate of the previous quarter, the non-zero LGD increases by around 0.6% on average). Our result supports recent studies in the literature of foreclosure spill-over effect, which shows strong evidence on negative impact of local foreclosure on house prices (e.g., Immergluck and Smith, 2006; Lin et al., 2009; Campbell et al., 2011; Anenberg and Kung, 2014; and Gerardi et al., 2015). An increase in the average local foreclosure rate amplifies the seller to buyer ratio giving more power to the buyer in the housing market. Hence, foreclosures may decrease the overall price and liquidity of the whole market (including both foreclosed and non-foreclosed properties)[15]. This process will worsen the collateral shortfall, and hence, increase the realised losses when the property is a forced sale. On another aspect of the housing market, we find evidence supporting that house price under distress of the foreclosed property exacerbates the loss severity. It is statistically estimated that a 10% increase in the house price under common market condition to house price under distress condition ratio in the previous quarter is associated with around 2.7% increase in the current non-zero loss severity on average. This is not surprising because foreclosed properties are subject to poor maintenance by previous owners, threats of vandalism as well as protection costs incurred by the mortgage lenders. These factors that lead to foreclosed properties are usually attached with favourable yet inevitable foreclosure discounts, which in turn may result in higher losses when they are realised.

5 Conclusion

This paper develops a three-step selection model with a joint probability framework for the probability of default, cure rate and magnitude of non-zero LGD for mortgage loans.

[15] In the recent work, Guren and McQuade (2015) have developed a dynamic search model in which this mechanism is revealed.

Statistically, our approach fits the bimodal distribution of LGD more efficiently, which shows a massive concentration on cured loans. Our empirical test supports a utilisation of dependence structure among default, cure and non-zero LGD (or equivalently, their joint probability framework) for the analysed sample. Interestingly, we find that the Dependent model works better than the Independent model and the current benchmark OLS model in the literature and industry in terms of out-of-time predictive accuracy. This difference is particularly remarkable for high-value ranges of LGDs during the GFC. Correlations among variables tend to be considerably higher during the crisis periods. Our model has important implications for bank risk models and prudential regulation not only because of its better predictive performance but also due to its suggestion in capital allocation. More specifically, our model provides a new approach to improve banks' efficiency in capital utilisation because capital costs of the expected losses will be meaningfully lower if a portion of defaulted loan in a bank's residential mortgage portfolio can be adequately linked with a high cure probability.

Our empirical findings on the US residential mortgage loans observed between Q1 2004 and Q1 2015 with regard to control variables are consistent with previous studies and economic intuition. Further, we provide new evidence that: (i) higher FICO loans, that are foreclosed, are less likely to be fully recovered; (ii) regional foreclosure contagion effect exists in the sense that higher local foreclosure rate increases the probability of default and worsens the non-zero LGD; and (iii) higher average gap between the local house price under common market condition and under distress condition exacerbates the non-zero LGDs. These findings suggest a necessity to include the FICO score to bank risk models due to its explanatory power to the likelihood that defaulted loans are cured. Further, an exploitation of the historical average local default rate and foreclosed house price discounts information also helps to explain the home loan losses.

References

Agarwal, S., Ambrose, B. W., Chomsisengphet, S., and Sanders, A. B. (2012). Thy neighbor's mortgage: Does living in a subprime neighborhood affect one's probability of default? *Real Estate Economics, 40*, 1-22.

Akkoc, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: the case of Turkish credit card data. *European Journal of Operational Research, 222*, 168-178.

Altman, E. I., Brady, B., Resti, A., and Sironi, A. (2005). The link between default and recovery rates: Theory, empirical evidence and implications. *Journal of Business, 78*, 2203-27.

Amromin, G., and Paulson, A. (2009). Comparing patterns of default among prime and sub-prime mortgages. *Economic Perspectives, 33*, 18-38.

Andersson, F., and Mayock, T. (2014). Loss severities on residential real estate debt during the Great Recession. *Journal of Banking and Finance, 46*, 266-284.

Anenberg, E., and Kung, E. (2014). Estimates of the size and source of price declines due to nearby foreclosures. *American Economic Review, 104*, 2527-51.

Bade, B., Roesch, D., and Scheule, H. (2011). Default and recovery risk dependencies in a simple credit risk model. *European Financial Management, 17*, 120-144.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*, 321-329.

Bajari, P., Chu, C. S., and Park, M. (2008). *An empirical model of subprime mortgage default from 2000 to 2007.* NBER Working Paper #14625.

Bellotti, T., and Crook, J. (2008). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society, 60*, 1699-1707.

Bellotti, T., and Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting, 28*, 171-182.

Bijak, K., and Thomas, L. C. (2015). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of the Operational Research Society, 66*, 342-352.

Boyes, W., Hoffman, D., and Low, S. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics, 40*, 3-14.

Calem, P. S., and LaCour-Little, M. (2004). Risk-based capital requirements for mortgage loans. *Journal of Banking and Finance, 28*, 647-672.

Campbell, J. Y., and Cocco, J. F. (2015). A model of mortgage default. *Journal of Finance, 70*, 1495-1554.

Campbell, J. Y., Giglio, S., and Pathak, P. (2011). Forced sales and house prices. *American Economic Review, 101*, 2108-31.

Chava, S., Stefanescu, C., and Turnbull, S. (2011). Modeling the loss distribution. *Management Science, 57*, 1267-1287.

Clauretie, T. M., and Herzog, T. (1990). The effect of state foreclosure laws on loan losses: evidence from the mortgage insurance industry. *Journal of Money, Credit, and Banking, 22*, 221-233.

Crawford, G. W., and Rosenblatt, E. (1995). Efficient mortgage default option exercise: Evidence from loss severity. *Journal of Real Estate Research, 10*, 543-555.

Crook, J., and Banasik, J. (2012). Forecasting and explaining aggregate consumer credit delinquency behaviour. *International Journal of Forecasting, 28*, 145-160.

Crook, J., and Bellotti, T. (2010). Time varying and dynamic moels for default risk in consumer loans. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*, 283-305.

Crook, J., Edelmann, D., and Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*, 1447-1465.

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

Demyanyk, Y., Koijen, R. S., and Van Hemert, O. A. (2011). *Determinants and consequences of mortgage default.* Federal Reserve Bank of Cleveland Working Paper.

Elul, R., Souleles, N., Chomsisengphet, S., Glennon, D., and Hunt, R. (2010). What "triggers" mortgage default? *American Economic Review, 100*, 490-494.

FICO. (2008). *Scoring your customers: How often is often enough?* Available at: http://www.fico.com/en/wp-content/secure_upload/Insights_Score_Migration_2505WP.pdf.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*, 368-378.

Frye, J. (2000a). Collateral damage. *Risk*, 91-94.

Frye, J. (2000b). *Collateral damage detected.* Working paper, Emerging Issues Series, Federal Reserve Bank of Chicago.

Gerardi, K., Herkenhoff, K. F., Ohanian , L. E., and Willen, P. S. (2013). *Unemployment, negative equity, and strategic default.* Federal Reserve Bank of Atlanta Working Paper.

Gerardi, K., Willen, P. S., Rosenblatt, E., and Yao, V. W. (2015). Foreclosure externalities: New evidence. *Journal of Urban Economics, 87*, 42-56.

Goodman, L. S., Ashworth, R., Landy, B., and Yin, K. (2010). Negative equity trumps unemployment in predicting defaults. *Journal of Fixed Income, 19*, 67-72.

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the World Economy, 10*, 299-316.

Guren, A. M., and McQuade, T. J. (2015). *How do foreclosures exacerbate housing downturns.* Working paper.

Immergluck, D., and Smith, G. (2006). The external costs of foreclosure: the impact of single-family mortgage foreclosures on property values. *Housing Policy Debate, 17*, 57-79.

Jarrow, R. A. (2001). Default parameter estimation using market prices. *Financial Analysts Journal, 5*, 75-92.

Jokivuolle, E., and Peura, S. (2003). Incorporating collateral value uncertainty in loss given default estimates and loan-to-value ratios. *European Financial Management, 9*, 299-314.

Lee, Y., Rösch, D., and Scheule, H. (2016). Accuracy of mortgage portfolio risk forecasts during financial crises. *European Journal of Operational Research, 249*, 440-456.

Lekkas, V., Quigley, J. M., and Van Order, R. (1993). Loan loss severity and optimal mortgage default. *Journal of American Real Estate and Urban Economics Association, 21*, 353-371.

Leow, M., and Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting, 28*, 183-195.

Leow, M., Mues, C., and Thomas, L. (2014). The economy and loss given default: evidence from two UK retail lending data sets. *Journal of the Operational Research Society, 65*, 363-375.

Li, X., Qi, M., and Zhao, X. (2012). *The extent of strategic defaults induced by mortgage modification programs.* OCC and York University Working Paper.

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

Loterman, G., Brown, I., Martens, D., Mues, C. and Baesens, B., (2012). Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting, 28(1), pp.161-170.

Maldonado, S., Pérez, J., and Bravo, C. (2017). Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research, 261*, 656-665.

Malik, M., and Thomas, L. (2009). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society, 61*, 411-420.

Matuszyk, A., Mues, C., and Thomas, L. C. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society, 61*, 393-398.

Pennington-Cross, A. (2003). *Subprime and prime mortgages: Loss distributions.* Working paper, Office of Federal Housing Enterprise Oversight.

Qi, M., and Yang, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance, 33*, 788-799.

Roesch, D., and Scheule, H. (2014). Forecasting probabilities of default and loss rates given default in the presence of selection. *Journal of the Operational Research Society, 65*, 393-407.

Johnston Ross, E.B. and Shibut, L., (2015). What Drives Loss Given Default? Evidence from Commercial Real Estate Loans at Failed Banks, FDIC CFR Working Paper 2015-03.

Jortzik, S. and Scheule, H., (2017). A Theoretical and Empirical Analysis of Alternative Discount Rate Concepts for Computing LGDs Using Historical Bank Workout Data. Available at: https://ssrn.com/abstract=2979430.

Seiler, M. J., Lane, M. A., and Harrison, D. M. (2014). Mimetic herding behavior and the decision to strategically default. *Journal of Real Estate Finance and Economics, 49*, 621-653.

Tong, E. N., Mues, C., and Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting, 29*, 548-562.

Tong, E., Mues, C., and Thomas, L. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218*, 132-139.

Wolter, M., and Rösch, D. (2014). Cure events in default prediction. *European Journal of Operational Research, 238*, 846-857.

Yao, X., Crook, J., and Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, Forthcoming, https://doi.org/10.1016/j.ejor.2017.05.017.

Zhang, Y., Ji, L., and Liu, F. (2010). *Local housing market cycle and Loss Given Default: Evidence from sub-prime residential mortgages.* Working paper, International Monetary Fund.

Zhang, Z., Gao, G., and Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. *European Journal of Operational Research, 237*, 335-348.

Appendix A – Derivation of the log likelihood function

Following the selection mechanism shown in Figure 2, we have three components of the likelihood function as follows:

1. *If the loan is not defaulted:*

$$\Pr(D_{it}=0)=\Pr\left(D_{it}^{*}\le 0\middle|X_{i,t-1}\right)=\Pr\left(X_{i,t-1}\beta+u_{it}\le 0\right)=\Pr\left(u_{it}\le -X_{i,t-1}\beta\right)=$$
$$\Phi\left(-X_{i,t-1}\beta\right)=1-\Phi\left(X_{i,t-1}\beta\right)$$

2. *If the loan is defaulted but it is cured:*

$$\begin{aligned}\Pr(D_{it}=1,C_{it}=1)&=\Pr\left(D_{it}^{*}>0\middle|X_{i,t-1},C_{it}^{*}>0\middle|\Theta_{i,t-1}\right)\\&=\Pr\left(u_{it}>-X_{i,t-1}\beta,v_{it}>-\Theta_{i,t-1}\lambda\right)\\&=\int_{-X_{i,t-1}\beta}^{\infty}\int_{-\Theta_{i,t-1}\lambda}^{\infty}f(u_{it},v_{it})dv_{it}u_{it}\\&=\int_{-X_{i,t-1}\beta}^{\infty}\int_{-\Theta_{i,t-1}\lambda}^{\infty}\phi_2(u_{it},v_{it},\rho_{uv})dv_{it}u_{it}\\&=1-\Phi_2\left(-X_{i,t-1}\beta,-\Theta_{i,t-1}\lambda,\rho_{uv}\right)=\Phi_2\left(X_{i,t-1}\beta,\Theta_{i,t-1}\lambda,\rho_{uv}\right)\end{aligned}$$

3. *If the loan is defaulted and non-cured:*

According to the Bayes rule, we have:

$$\Pr\left(L_{it},D_{it}=1,C_{it}=0\middle|X_{i,t-1},\Theta_{i,t-1},Z_{i,t-1}\right)$$
$$=f\left(L_{it}\middle|Z_{i,t-1}\right)\Pr\left(D_{it}=1,C_{it}=0\middle|L_{it},X_{i,t-1},\varTheta_{i,t-1},Z_{i,t-1}\right)\qquad(12)$$

In (12), following density function of normal distribution it is easy to see that,

$$\begin{aligned}f\left(L_{it}\middle|Z_{i,t-1}\right)&=f(\varepsilon_{it})\\&=\frac{1}{\sigma_\varepsilon}\phi\left(\frac{L_{it}-Z_{i,t-1}\alpha}{\sigma_\varepsilon}\right)\end{aligned}$$

The remaining part of (12) can be derived as follows:

$$\begin{aligned}\Pr\left(D_{it}=1,C_{it}=0\middle|L_{it},X_{i,t-1},\varTheta_{i,t-1},Z_{i,t-1}\right)&=\Pr\left(D_{it}=1,C_{it}=0\middle|\varepsilon_{it},X_{i,t-1},\Theta_{i,t-1}\right)\\&=\Pr\left(D_{it}^{*}>0,C_{it}^{*}\le 0\middle|\varepsilon_{it},X_{i,t-1},\Theta_{i,t-1}\right)\\&=\Pr(-u_{it}<X_{i,t-1}\beta,v_{it}\le-\Theta_{i,t-1}\lambda|\varepsilon_{it})\qquad(13)\end{aligned}$$

The conditional distribution of $(-u_{it},v_{it})'$ given ε_{it} can be written as:

$$\begin{pmatrix}-u_{it}\\v_{it}\end{pmatrix}|\varepsilon_{it}\sim N\left[\begin{pmatrix}\frac{-\rho_{u\varepsilon}}{\sigma_\varepsilon}(L_{it}-Z_{i,t-1}\alpha)\\\frac{\rho_{v\varepsilon}}{\sigma_\varepsilon}(L_{it}-Z_{i,t-1}\alpha)\end{pmatrix},\begin{pmatrix}1-\rho_{u\varepsilon}^2&-\rho_{uv}+\rho_{u\varepsilon}\rho_{v\varepsilon}\\-\rho_{uv}+\rho_{u\varepsilon}\rho_{v\varepsilon}&1-\rho_{v\varepsilon}^2\end{pmatrix}\right]$$

Standardizing the bivariate conditional distribution, we have:

$$\begin{pmatrix} u_{it}^* \\ v_{it}^* \end{pmatrix} |\varepsilon_{it} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{uv|\varepsilon}^* \\ \rho_{uv|\varepsilon}^* & 1 \end{pmatrix}\right]$$

where,

$$u_{it}^* = \frac{-u_{it} + \frac{\rho_{u\varepsilon}}{\sigma_\varepsilon}(L_{it} - Z_{i,t-1}\alpha)}{\sqrt{1-\rho_{u\varepsilon}^2}}$$

$$v_{it}^* = \frac{v_{it} - \frac{\rho_{v\varepsilon}}{\sigma_\varepsilon}(L_{it} - Z_{i,t-1}\alpha)}{\sqrt{1-\rho_{v\varepsilon}^2}}$$

$$\rho_{uv|\varepsilon}^* = \frac{-\rho_{uv} + \rho_{u\varepsilon}\rho_{v\varepsilon}}{\sqrt{1-\rho_{u\varepsilon}^2}\sqrt{1-\rho_{v\varepsilon}^2}}$$

Hence, (13) is equivalent with,\ :

$$\Pr\left(u_{it}^* < \frac{X_{i,t-1}\beta + \frac{\rho_{u\varepsilon}}{\sigma_\varepsilon}(L_{it} - Z_{i,t-1}\alpha)}{\sqrt{1-\rho_{u\varepsilon}^2}}, v_{it}^* \le \frac{-\Theta_{i,t-1}\lambda - \frac{\rho_{v\varepsilon}}{\sigma_\varepsilon}\left(L_{it} - Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{v\varepsilon}^2}} |\varepsilon_{it}\right)$$
$$= \Phi_2\left(\frac{X_{i,t-1}\beta + \frac{\rho_{u\varepsilon}}{\sigma_\varepsilon}(L_{it} - Z_{i,t-1}\alpha)}{\sqrt{1-\rho_{u\varepsilon}^2}}, \frac{-\Theta_{i,t-1}\lambda - \frac{\rho_{v\varepsilon}}{\sigma_\varepsilon}\left(L_{it} - Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{v\varepsilon}^2}}, \rho_{uv|\varepsilon}^*\right)$$

Combining the components of likelihood function from the three components we have the full likelihood function of the model as:

$$L = \prod_{t=1}^{T}\prod_{i=1}^{N}\left(1-\Phi\left(X_{i,t-1}\beta\right)\right)^{1-D_{it}} \cdot \left(\Phi_2\left(X_{i,t-1}\beta, \Theta_{i,t-1}\lambda, \rho_{uv}\right)\right)^{D_{it}\cdot C_{it}}$$
$$\cdot\left(\frac{1}{\sigma_\varepsilon}\phi\left(\frac{L_{it} - Z_{i,t-1}\alpha}{\sigma_\varepsilon}\right)\right.$$
$$\left.\cdot\,\Phi_2\left(\frac{X_{i,t-1}\beta + \frac{\rho_{u\varepsilon}}{\sigma_\varepsilon}\left(L_{it} - Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{u\varepsilon}^2}}, \frac{-\Theta_{i,t-1}\lambda - \frac{\rho_{v\varepsilon}}{\sigma_\varepsilon}\left(L_{it} - Z_{i,t-1}\alpha\right)}{\sqrt{1-\rho_{v\varepsilon}^2}}, \rho_{uv|\varepsilon}^*\right)\right)^{D_{it}\cdot(1-C_{it})}$$