Modelling prepayment risk in residential mortgages

Janneke Meis
372866

Msc Thesis Quantitative Finance
ERASMUS UNIVERSITY ROTTERDAM
November 2015

Supervisors:
Dr. Rogier Potter van Loon (Erasmus)
Maurits Malkus (Deloitte)
Michiel Hopman (Deloitte)
Co-reader:
Dr. Marcin Jaskowski

Abstract

The current low interest rate environment has triggered refinancing incentives in the residential mortgage sector. An unscheduled return of a part or the full outstanding principal constitutes a risk from the perspective of financial institutions providing mortgages. On the one hand, prepayments affect the Asset and Liability management of a bank. On the other hand, prepayments lead to interest rate risk. Given the magnitude of the residential mortgages on the balance sheet of a bank, it is of vital importance to obtain insight in the actual maturity of the mortgages provided by financial institutions. This thesis looks into different models that can be used to predict current and future prepayment rates. The most widely used prepayment models are option theoretic models, multinomial logit models and competing risk models. Aside from these models this thesis investigates the applicability of the Markov model as a prepayment model. Important determinants of prepayment include borrower specific characteristics, loan specific characteristics and macro-economic variables. Variable selection procedures are used to identify the most important risk drivers. Models are compared based on a wide range of in-sample and out-of-sample performance criteria to determine the model that is most appropriate for predicting prepayment rates. Another important feature that the models should be capable of incorporating is the recent credit crisis of 2008. Tests on parameter stability are conducted to determine the possible presence of structural breaks in prepayment models.

Contents

1 Introduction

The outstanding amount of all residential mortgage loans in the US amounted USD 13.5 trillion in the beginning of 2015[1]. A total of USD 7.6 trillion has been securitized and sold to the secondary market in the form of mortgage backed securities (MBS) and mortgage trusts. The remainder of the outstanding debt constitutes a direct investment in mortgage loans by financial institutions. The most important segment in the secondary mortgage market are the so-called agency mortgage backed securities. The three agencies that issue and guarantee mortgages in the US are the Government National mortgage Association (Ginnie Mae, GNMA), the Federal National Mortgage Association (Fannie Mae, FNMA) and the Federal Home Loan Mortgage Corporation (Freddie Mac, FHLMC). The combined outstanding debt held by these agencies amounted USD 2.7 trillion at the beginning of 2015. GNMA securities are default free since they are fully backed by the US government. FNMA and FHLMC are government sponsored agencies but are not default free. GNMA, FNMA and FHLMC securities generally have highly standardized features, trading and settle mechanisms. The most common mortgage in their portfolio is the thirty year fixed rate fully amortizing mortgage.

Standard residential mortgages in the US offer full prepayment flexibility in the sense that the sense that the entire principal outstanding can be paid off at any time. Unlike in the Netherlands the majority of mortgages can be prepaid without a prepayment penalty. The option to prepay can be seen as an American option on the mortgage contract. At the discretion of the borrower, the outstanding principal can be returned, relieving the mortgagee from any future contractual obligations on the loan. Allthough it is possible to form an estimate of the timing of option exercise, such options are often not exercised in an optimal way. Main causes are the diverging ability of borrowers to time optimal exercise of such options as well as other non-rational heterogeneities in borrower behaviour. Aside from the option to prepay the borrower also has the 'option' to default on the mortgage. The incentive for voluntary default is high when the market value of the property is lower than the value of the mortgage. Again, although this option may be exercised optimally, heterogeneity in borrower behaviour prevents the prediction of option exercise times in a purely optimal fashion.

The presence of implicit options embedded in mortgage contracts coupled with heterogeneity in borrower behaviour pose risks from the side of the mortgagor. On the one hand, uncertainty related to predicting the timing of cash flows leads to liquidity risk. One other hand, a mismatch in interest rates paid and received leads to interest rate risk. From the perspective of the lender it is therefore important to obtain reliable estimates of future prepayment and default rates as well as gain insight in the factors driving these rates. This can aid financial institutions in their Asset and Liability management.

The most important determinant of prepayment is refinancing. Refinancing is attractive when the mortgage rate currently paid on the loan is higher than the mortgage rate available in the market. Other reasons for prepayment are for example relocation and divorce. Default is a form of prepayment in which the outstanding principal is returned via a sale of the property. A peculiar feature of the US mortgage market is that a borrower with mortgage (pre-)arrears can opt for a voluntary repossession. In some states in the US, borrowers can hand in the keys of their home and be relieved from any future obligations on the mortgage.

The US economy is slowly recovering from the collapse of the housing and mortgage markets in 2008. House prices are rising again since their bottom out in 2011. Home sales remain generally flat but housing starts are again increasing. Since the decrease in mortgage rates as of 2013 not only has the demand for mortgage loans increased, the level of prepayments has dropped after its

[1] Source: Economic Research Data of the Board of Governors of the Federal Reserve System, available at http://www.federalreserve.gov/econresdata/releases/mortoutstand/current.htm.

spike in 2009. Mortgage defaults are on a decreasing trend since the crisis (Agency, 2014). All in all, current market developments lead to rising fear among lenders for increasing prepayment rates. The housing market is picking up which fuels incentives to move. Simultaneously, although interest rates are beginning to rise, they still remain at a historically low level and are expected to rise in the future. Searching for better mortgage rates in the market is likely to be a fruitful action.

Given the size and the developments of the US mortgage market in the last decade this market is an interesting market to study. A prepayment model developed now is likely to be considerably different from the prepayment models constructed a decade ago. Aside from being able to capture the economic conditions of the last decade, a model for prepayments should contain both an optimal component as well as a component that incorporates borrower heterogeneity. The aim of the present research is to develop a model that is able to accurately predict past, current and future prepayment rates in the US. This involves the selection of appropriate modelling strategies as well as the selection of relevant variables. Furthermore, attention will be paid to the changing circumstances resulting from the business cycle as well as borrower heterogeneity.

This research uses the Single Family Loan Level data set provided by Freddie Mac. It contains information on mortgage characteristics as well as mortgage performance over time for loans issued by Freddie Mac in the period 1999 to 2014. Additionally, several macro-economic variables are added to the data set. Prepayments are modelled with an option theoretic model, a multinomial logit (MNL) model, a competing risk model and a Markov model. The MNL model is the most widely used in the research on mortgage termination due to its relatively straightforward estimation. It is demonstrated that the competing risk model constitutes a special case of the MNL model under certain conditions. The Markov models are estimated as conditional logit models using covariate information and conditioning on the current status of contractual mortgage payments. Due to the multitude of variables affecting prepayment and default decisions, a variable selection procedure is applied to determine which variables are most influential for explaining mortgage termination. Since it has been established that prepayments exhibit a behavioural uncertainty, the results of the option theoretic model are not of interest standing alone but are included as an explanatory variable in the exogenous models. The optimal refinancing incentive is derived via the risk neutral valuation of a European lookback put option on the mortgage contract. To this end, the risk free rate is modelled with a Hull-White One Factor (HW1f) model. The advantage of this model is that it can provide an exact fit to the current term structure by taking this as an additional input variable, next to the spot rate. Since refinancing constitutes the most important determinant for prepayments, the appropriate construction of the refinancing incentive is a valuable addition to the exogenous models.

The aim of this thesis is to develop a prepayment model that can predict in-sample and out-of-sample prepayment rates, whereby out of sample can be cross sectional as well as time series wise. Performance is assessed by means of a wide array of indicators. Models under investigation are the prepayment models most widely used in the prepayment literature. The Markov model is introduced as a novel prepayment model. A multitude of loan specific, borrower specific and macro economic variables are included in the analysis as risk drivers. Special attention is paid to whether the models hold up well during the financial crisis of 2008. Prepayments and defaults cause the actual maturity of a mortgage contract to be stochastic and typically shorter than its contractual maturity. The main contribution of this thesis is to provide insight in the true maturity of a portfolio of mortgage contracts. This is relevant for financial institutions since the funding they attract to cover for the mortgages they provide should be of shorter maturity as well. This research adds to existing literature on this topic by applying the most common models in the prepayment literature as well as a novel prepayment model, the Markov model, to one data set. This allows for an accurate comparison between the models for which a wide array of performance criteria will be used. Moreover this research pays special attention to the presence of a structural break since the data set includes a the 2008 credit crisis.

The main findings of this thesis are the following. Variable selection procedures for the different models revealed that the most important determinants for prepayment are loan size, the unemployment rate, the Loan-to-Value ratio, housing prices and the refinancing incentive. Furthermore, jointly estimating prepayments and defaults is beneficial for the model. However, including intermediate states such as partial prepayment and delinquency status does not lead to improved model performance in the MNL models. The derivation of the lookback put option is interesting for pricing embedded options in the mortgage contract into the mortgage rate paid by borrowers, but its contribution to prepayment estimation is limited. Its limitations mainly lie in relatively high computation time involved in deriving the value of the put. A refinancing incentive based on contractual and market mortgage rate differential as well as outstanding contractual payments is sufficient to capture this most important determinant for prepayments. The models perform very well in determining the mortgage characteristics relevant for predicting prepayments. Cross sectional model performance is assessed based on contingency tables. Time series wise, the models suffer from some autocorrelation in the residuals however the estimates of the models are relatively accurate. Although information on the current state is Incorporated in the Markov models, this is insufficient for capturing dependence over time. Based on the selected set for performance indicators, the three state multinomial logit model is the selected as best performer. Finally, a structural break in prepayment rates is found shortly after the crisis. This is incorporated by allowing for different parameters pre and post crisis.

The remainder of this thesis is structured as follows. Section 2 will provide a general background on the US mortgage market, including the regulations concerning prepayment and default. Different variables affecting mortgage termination will be discussed. In Section 3 the technical background for the option theoretic model, the MNL model, the competing risk model and Markov model will be given. Section 4 contains the data description and includes bivariate analyses of a wide set of risk drivers. Section 5 discusses the estimation outputs of the models and presents the results on cross sectional and time series out-of-sample forecasts. Section 6 compares the models based on a set of performance criteria and tests for structural breaks in the model that is selected as best performer. Section 7 concludes.

2 General background

2.1 Mortgage market in the United States

The US mortgage market is characterized by a relatively high negative equity rate. Although the rate has come down from 31.4 percent in the second quarter of 2012 to 15,4 percent in the first quarter of 2015, this percentage is still high in international comparison (Gudell, 2015). In 2015Q1, 11.8 percent of home owners in negative equity owe more than twice what their home is worth to the bank. Even though such high loan-to-value (LTV) ratios[2] are geographically concentrated, the US mortgage market can in general be characterized as having high LTV ratios. Mortgage defaults are on a downward trend as well, although still well above pre-crisis levels. Although mortgages rates have risen slightly since 2012, they still remain at relatively low levels, increasing the propensity to refinance (Agency, 2014).

The most common mortgage product in the US is the fixed rate fully amortizing loan with a maturity of thirty years. The rate paid on the mortgage is generally fixed for the lifetime of the loan and the the loan is fully paid off at maturity. Each period, a constant amount is repaid. The fraction of this amount that is used to repay the principal increases over time since interest payments decrease as the outstanding balance decreases over time.

An important feature of the US residential mortgage market is that agency backed loans allow for a penalty free prepayment of (part of) the principal any time before maturity. Commercial and investment property loans can however carry a prepayment penalty, which often consists of a percentage of the remaining principal outstanding. To discourage prepayment behaviour, financial institutions in other countries often set a monthly prepayment limit in excess of which a penalty will be charged.

The default legislation in the US is often targeted as one of the causes of the financial crisis. The most peculiar feature is the fact that in some states[3] mortgage loans are nonrecourse, meaning that if a borrower fails to make contractual mortgage payments, the lender can seize the collateral but has no recourse to any other of the borrower's assets (Gerardi and Hudson, 2010). Voluntary repossession effectively means that a borrower can hand in the keys of the house an be free from any future obligations even though he has enough cash on the bank. Especially in times when housing prices are low, the existence of negative equity leads to a large amount of so-called strategic defaults. In most countries, the mortgage loan is diminished by the amount paid at the auction. The difference between the market value of the property and the value of the loan, the shortfall, is added to the debt of the borrower. To be alleviated from the entire mortgage debt, borrowers have to file for bankruptcy.

Voluntary repossessions are not entirely free of costs. They come at the expense of a downgrade in the FICO credit score. FICO scores are used by many US banks to assess the creditworthiness of individuals. The score is designed to measure the risk of default by taking into account various factors in a person's financial history, such as payment history, size of the debt burden and length of credit history. Lenders are in the position to offer different terms and conditions to customers depending on their credit rating.

[2] Loan-to-value is the ratio of current loans divided by indexed house values.

[3] The following states are classified as nonrecourse states: Alaska, Arizona, California, Iowa, Minnesota, Montana, North Carolina, North Dakota, Oregon, Washington and Wisconsin (Gerardi and Hudson, 2010).

2.2 Embedded options in the mortgage contract

The mortgages sold by mortgagees contain implicit options. Prepayment is one of these options. The mortgagor exchanges the unpaid principal balance on the mortgage for a release from further obligations (Deng, 1997). Default is another optionality in a mortgage contact. The property is sold in exchange for elimination of future mortgage obligations. The difficulty with quantifying such options is that mortgagors do not exercise these options efficiently. Moreover, mortgagors' behaviour is heterogeneous and cannot be represented by a typical borrower (Deng et al., 2005). This risk that the behaviour of the mortgagor deviates from what is expected from a purely financial standpoint is called behavioural risk (Bissiri and Cogo, 2014). The application of this risk to early unscheduled return of the principal on a mortgage is defined as prepayment risk. A formal definition is given in Kolbe (2002) who defines prepayments as: '(contractually permitted) notional cash flows which occur earlier or later than expected, deviating form the anticipated call or put policy of the counterparty in a financial contract' (Kolbe, 2008), p.21. Henceforth, Kolbe defines prepayment risk as the risk resulting from these cash flow deviations.

Risk of early mortgage termination makes the duration of a portfolio of mortgages stochastic and in turn has implications on the refinancing policy of the lender (Jacobs et al., 2005). The implicit options embedded in mortgages may or may not be exercised in response to market changes, which in turn leads to significant liquidity risk and interest rate risk for the credit providing institution (Consalvi and di Freca, 2010). Interest rate risk in general arises when there is a mismatch in the fixation of the interest rates paid and received by the bank (Perry et al., 2001). To fund the mortgage, financial institutions attract resources from elsewhere, on which a certain agreed upon rate has to be paid over a fixed period. If the mortgagee decides to prepay (part of) the principal at any time before its original maturity, the mortgagee will have to find an alternative use for these funds. If market interest rates have fallen since the origination of the mortgage the bank will incur a loss. On the other hand, uncertainty in the maturity profile of loans subjective to prepayment has a considerable impact on the representation of the liquidity profile of the bank (Consalvi and di Freca, 2010). An incorrect evaluation of this profile exposes financial institutions to the risk of overestimating future liquidity requirements (overfunding) as well as to the risk of increased long-term liquidity costs (Bissiri and Cogo, 2014).

To incorporate the risk of prepayment, financial institutions generally include a charge in their mortgage pricing. This risk is however not priced in completely as a mortgagor can prepay any time while the spread to account for this risk is received on a monthly basis across the life of a mortgage (Vasconcelos, 2010). To discourage prepayment behaviour of mortgagees, banks usually charge a prepayment penalty. This penalty is mostly equal to the present value of the difference in monthly interest payments between the mortgage and of a newly originated mortgage with the same characteristics (Jacobs et al., 2005).

2.3 Determinants of mortgage termination

There exist a multitude of factors that influence the mortgagors' decision of mortgage termination. These risk drivers can roughly be divided into borrower-specific factors, loan-specific factors and macro-economic factors, see Table 1. The effect of these variables on prepayment and default rates is fairly straightforward, see for example (Clapp et al., 2000).

Some important features stand out. Firstly, prepayments often exhibit an S-shaped relation with loan age. This so-called seasoning effect arises since prepayment rates are generally low shortly after origination of the loan and increase as the mortgage matures, the ramp-up period, to finally arrive at a steady state level near the maturity of the mortgage (Charlier and Bussel, 2001), (Jacobs et al., 2005). Secondly, trends in house prices are indicative of the level of activity in the housing and

Category	**Explanatory variables**
Borrower specific	Age, income, creditworthiness, loan purpose, employment status.
Loan specific	Loan age, loan amount, mortgage rate, insurance, penalty, location, property type, market value of the property.
Macro-economic	Housing prices, mortgage rates, risk free rates, divorce rate, month of the year.

Table 1: Factors affecting prepayment rates.

mortgage markets. In periods of house price appreciation, home sales and mortgage originations may increase as the expected return on investment rises (Agency, 2014). Prepayments due to relocation are more prevailing in this case. Conversely, during periods of price depreciation or price uncertainty, home sales and mortgage originations tend to decrease as risk-averse home-buyers are reluctant to enter the market. In turn, this leads to fewer prepayments due to relocation. Furthermore, prepayment due to relocation is positively affected by the divorce rate.

Prepayment due to refinancing incentives is by far the most important reason for prepayment. Refinancing can be attractive when the mortgage market rate is below the contractual rate. The contractual mortgage rate paid and the effective duration of the mortgage are inversely related (Burns, 2010). An important feature of the refinancing incentive is the so-called burnout effect (Jacobs et al., 2005). This effect arises due to differences in borrower behaviour in a pool of mortgages. If a refinancing incentive occurs (such as a drop in mortgage market rates) a wave of prepayments will occur. The borrowers that grasp this opportunity can be deemed the 'fast',e.g. financially aware, borrowers whereas the remainder of the borrowers in the pool are 'slow' borrowers. Therefore, in the presence of another refinancing opportunity, the pool is expected to be less active or in other words, the pool is burned out (Gonchanov, 2002).

Going into strategic default when the market value of the property is lower than the value of the loan is the second most important determinant for prepayment. This is especially true in jurisdictions where mortgage loans are issued on a non-recourse basis. Properly assessing the value of negative equity at each point in time is however a challenging task since the market value of the property is unknown.

2.4 Prepayment models

Following on the previous section, it is clear that it is challenge to model prepayment rates give the multitude of factors that are of influence in this decision. The models proposed in the literature can broadly be classified into optimal prepayment models on the one hand and exogenous prepayment models on the other hand. The former category contains models that consider prepayments to occur as a result of fully rational behaviour of borrowers. These models make the assumption that prepayments are exercised in an optimal way and rely on the absence of arbitrage. Only taking financial considerations into account would lead mortgagors to prepay if the current value of their property exceeds the remainder of the outstanding mortgage plus transaction costs (Bussel, 1998). An entire body of literature has developed which models prepayments as options on a mortgage. This assumptions is useful for valuation purposes but is quite stringent in ruling out all irrational behaviour. Studies have demonstrated that modelling prepayments according to financial variables such as housing turnover and mortgages rates can lead to either underestimation of prepayment rates under financial optimal circumstances and overestimation of prepayment rates in financial suboptimal times (Consalvi and di Freca, 2010). Indeed, it has been observed that prepayments contain a considerable stochastic component following behavioural uncertainty.

Studies have responded by modelling this stochastic component in prepayments separately from the 'optimal' component. This lead to the development of exogenous models, which model prepayments using different explanatory variables. The aim of these models is to explain the relationship between observed prepayment rates and a set of explanatory variables. This influence can be assessed on a pooled level and on a loan level, if such data are available. Borrower heterogeneity can then be accounted for by including borrower characteristics in these models. If data on individual borrowers is unavailable, differences in the behaviour of borrowers can be modelled by incorporating the possibility that different types of borrowers exist (Quigley et al., 2000). In this way, unobserved borrower heterogeneity, such as borrower tastes and abilities, can be incorporated.

2.5 Prepayment modelling over time

The prepayment projections resulting from the models are dependent on the time frame in which they are conducted. Prior to the credit crisis of 2008 a prepayment model might predict a sharp increase in prepayment level, following a refinancing incentive. However in 2008 and 2009 a similar prepayment model should predict a lower prepayment rate following the same refinancing incentive (Burns, 2010). Economic circumstances have considerable influence on prepayment rates. More specifically, a weakened housing market, unemployment levels and lending standards create a 'liquidity crisis' for mortgagors and thereby reduce the incentive to prepay. Therefore, any prepayment model should contain macro-economic factors to account for economic conditions in a country. Furthermore, it is interesting to test for structural breaks in the data set. It is likely that parameter estimates prior and post 2008 are considerably different.

3 Technical background

This section provides a technical background on two different methods for modelling mortgage termination: option theoretic models and empirical (exogenous) models.

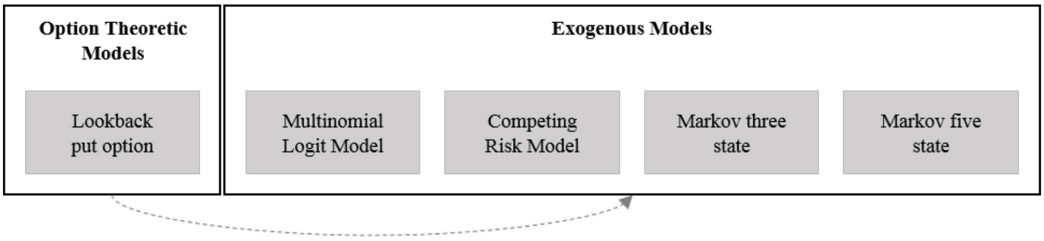


Figure 1: Overview of prepayment models.

Option theoretic models assume that borrowers behave purely rational and model prepayments endogenously, without the use of explanatory variables. Empirical models on the other hand model mortgage termination exogenously and can incorporate borrower heterogeneity. Comparing the two streams in modelling is indicative for the significance of behavioural risk in models for mortgage termination. The two models can be combined by including the result of the option theoretic model as an explanatory variable in the exogenous models. The following sections will discuss the Multinomial Logit Model (MNL), competing risk model, Markov model and an option theoretic model in turn.

3.1 Empirical models

3.1.1 Multinomial logit model

Discrete choice models are models in which the dependent variable is a categorical response variable, $Y_{it} = j$ for $j = 1, 2, ...$ Using a set of explanatory variables, these models estimate the probability that that either one of the categories in the ordinal dependent variable occurs $P(Y_{it} = j)$. Two popular models are the probit model on the one hand and the logit model on the other hand. The difference between the two lies in the distributional assumptions for the error terms. The probit model assumes a normal distribution, whereas the logit model assumes a logistic distribution. In the context of mortgage continuation and termination, the multinomial logit model (MNL) is the most widely adopted. The logit is defined as

$$\ln\left(\frac{F(.)}{1-F(.)}\right) = X_{it}'\beta_j, \tag{3.1}$$

where $F(.)$ is the logistic cumulative distribution function, X_{it} contains the explanatory variables and β_j denote coefficient estimates. Consequently, let the probability that mortgage i at time t is classified as category j be defined as

$$P(Y_{it} = j) = \frac{e^{X_{it}\beta_j}}{\sum\limits_j e^{X_{it}'\beta_j}}. \tag{3.2}$$

Competing risks are included in the MNL model through the restriction that the probabilities of the categories must sum to one, $\sum\limits_j P(Y_{it} = j) = 1$. Hence, a probability increase for one of the categories must necessarily be associated with a probability decrease for one of the other categories. The coefficients in the MNL model are interpreted in terms of the log odds ratio. To ensure parameter identification, one of the categories is set as benchmark category. To this end, the coefficients are equated to zero. The probability of the i'th mortgage at time t being classified as being in the reference state follows directly from 3.2 and reads

$$P(Y_{it} = 0) = \frac{1}{1 + \sum\limits_{j-1} e^{X'_{it}\beta_j}}. \tag{3.3}$$

The log likelihood of the MNL model is defined as

$$\ln L(\beta) = \sum_i \sum_t \sum_j d_{ijt} \ln P(Y_{it} = j), \tag{3.4}$$

in which d_{ijt} is a dummy variable for the category of Y_{it}. The coefficients are estimated by means of Maximimum Likelihood Estimation (MLE) and are obtained from the first order conditions (FOC) of Equation (3.4). Estimates are obtained by maximizing $\ln L$ with respect to β_j.

One of the properties of the MNL model is the Independence of Irrelevant Alternatives (IIA). This means that the odds ratio for any pair of choices is assumed to be independent of any other alternative. Elimination of one of the choices should not change the ratios of probabilities for the remaining choices. This means that a requirement for the categories is that they are mutually exclusive. More formally, the IIA property states that characteristics of one particular choice alternative do not impact the relative probabilities of choosing other alternatives. To validate whether the IIA property holds the Haussman-McFadden test can be performed. This test relies on the insight that under IIA, the parameters of the choice under a subset of alternatives may be estimated with a MNL model on just this subset or the full set though the former is less efficient than the latter (Vijverberg, 2011). The test statistic is given by

$$HM = (\hat{\beta}_U - \hat{\beta}_R)'[\hat{V}_U - \hat{V}_R]^{-1}(\hat{\beta}_U - \hat{\beta}_R), \tag{3.5}$$

which is $\sim\ \chi^2(k)$ with k denoting the number of parameters. Under the null hypothesis IIA holds which implicates that omitting irrelevant alternatives will lead to consistent and efficient parameter estimates for the restricted model, $\hat{\beta}_R$, while the parameter estimates of the unrestricted model,$\hat{\beta}_U$ are consistent but not efficient. Under the alternative, only $\hat{\beta}_U$ are consistent.

In this thesis, two classifications for Y_{it} will be adopted. The three state classification includes contractual payment, default and prepayment

$$Y_{it}^3 = \begin{cases} 1 & \text{if Contractual payment} \\ 2 & \text{if Prepayment} \\ 3 & \text{if Default} \end{cases},$$

in which the first category will be taken as reference state.

Information on intermediate states states can be included via curtailments (partial prepayments) and delinquency (delayed payments). This leads to a dependent variable with five states

$$Y_{it}^5 = \begin{cases} 1 & \text{if Contractual payment} \\ 2 & \text{if Curtailment} \\ 3 & \text{if Prepayment} \\ 4 & \text{if Delinquent} \\ 5 & \text{if Default} \end{cases}.$$

The superscript will be omitted when the analysis is applicable to both models. The MNL constitutes the most common approach to modelling mortgage termination. Due to the relative ease with which such models can be estimated, they form an appropriate starting point for modelling mortgage termination. DCM suffer from the drawback that they are unable to take into account dependence in observations that arises from the fact that the same mortgage contract is observed over multiple time periods. Mortgage termination cannot be assumed to be independent over time. Survival models can alleviate this problem.

3.1.2 Survival analysis

Survival models specify the probability distribution for the duration of a mortgage contract. The dependence of observations over time is taken into account by using the mortgage contract as unit of measurement as opposed to the contract year.

Survival models focus on modelling the time until the occurrence of a certain event. Within the survival analysis the link between the survival function and the covariates is usually expressed on the basis of two models: accelerated life models (ALM) and proportional hazard models (PHM) (Consalvi and di Freca, 2010). Cox (1972) was among the first to model time to failure as an underlying random variable. After (Green and Shoven, 1983) applied survival models to the mortgage literature, this approach became more popular in this field of research (Charlier and Bussel, 2001), (Jacobs et al., 2005), (Deng et al., 2005). In these studies the duration of a mortgage is modelled until it is terminated. The hazard rate is defined as the probability of mortgage termination for reason j at time t given the non-occurrence of this event until time t

$$\lambda_j(t,x) = \lim_{\Delta t \to 0} \frac{P(t \leq T_j < t + \Delta t | T_j \geq t)}{\Delta t} = \lambda_{j0}(t) e^{-X'_{it}\beta_j}. \tag{3.6}$$

As can be seen, to analyze the relationship between the survival function and the covariates, the model can be split into two parts. The first part of the model, $\lambda_{0j}(t)$, captures the distribution of the failure time when the explanatory variables are equal to zero, e.g. the base rate. It is a reflection of the 'natural' prepayment rate and varies with the age of the mortgage. The seasoning effect explained in Section 2.3 is directly visible in this rate. The parametric specifications assume a specific functional form, such as Exponential, Weibull, Log Normal or Log logistic. These distributions are often selected due to their ability of capturing the S-shaped relationship between prepayments and age of the loan. Alternatively, the distribution can be left unspecified or be estimated non-parametrically. The second part of the model incorporates the effect of the explanatory variables on the hazard rate. The second part in (3.6) is a proportionality factor that incorporates both loan -and time specific effects, X_{ijt}.

Following (Rodriguez, 2005), let T be a discrete random variable representing survival time. Analogous to the classification of the dependent variable in the multinomial logit model, it is assumed that mortgage termination can occur due to prepayment or default. The default category is contractual payment. Since the data collection period is cut-off at a certain date, some mortgage duration data will be unobservable if these mortgages are not terminated before the cut-off date. These mortgages are generally classified as belonging to the benchmark category (contractual payment). The survival function is defined as the probability of surviving from failure type j up to time t and is given by

$$S_j(t,x) = e^{-\Lambda_j(t,x)}, \tag{3.7}$$

where $\Lambda_j(t,x)$ is the cumulative hazard for cause j

$$\Lambda_j(t,x) = \sum_t \sum_i \lambda_j(t,x). \tag{3.8}$$

Combining the above, the unconditional probability that a mortgage is terminated at time t due to cause j is given by the cause-specific density

$$f_j(t,x) = \lambda_j(t,x)S(t,x). \tag{3.9}$$

The log likelihood of the competing risk model is given by

$$\ln L(\beta_j) = \sum_t \sum_i \sum_j d_{ijt} \ln f_j(t,x), \tag{3.10}$$

where d_{ijt} is a dummy variable indicating the reason for contract termination.

This likelihood is very similar to the likelihood of the MNL model, see Appendix A. In a competing risk model the analysis can thus be broken down into two parts. The MNL model determines the cause of death and the standard hazard model determines the overall risk (Rodriguez, 2005). When the effect of the baseline hazard, λ_{0j} is small relative to the effect of the covariates the competing risk model and the MNL model are comparable.

One drawback of competing risk models is the difficulty of including time varying covariates in the analysis. This can be seen as follows. The dependent variable in survival studies constitutes the survival time of mortgage i from cause j, T_{ij}. This variable does not depend on time t over which mortgages are observed. Using time varying covariate information in competing risk models poses a problem since for each mortgage only one observation on each covariate can be used. Therefore, the panel dataset has to be transformed into a cross-sectional dataset by combining information per time-varying covariate per mortgage i. There are several methods for combining time varying information into one observation per covariate. One example given in Section 3.5.2 is to average time varying covariates per mortgage. Another option is to randomly select a time period t for each mortgage i from which covariate information is included.

Since it has been demonstrated in Appendix A that the MNL model and the competing risk model are very similar and given the fact that the MNL model is capable of including time varying covariates, the MNL model is preferred over the competing risk model. Therefore, the coefficients of the competing risk model are not estimated separately. The baseline hazard of the survival models will be investigated to determine whether the outcome of the competing risk model and MNL model is indeed similar.

A more appropriate model to account for dependence between observations is the Markov model. Dependence is accounted for by conditioning on the current state in mortgage termination. In three state models the preceding state is always contractual payment. In the five state models conditioning on partial prepayment and delinquency status can contribute to capturing dependency in prepayments and defaults.

3.1.3 Markov model

In general, a stochastic process $Y_t, t \geq 0$ with state space $\mathcal{S}$ is a discrete time Markov chain if, for all states $i, j, s_0, ..., s_{t-1} \in \mathcal{S}$,

$$P(Y_{t+1} = j | Y_t = i, Y_{t-1} = s_{t-1}, ..., Y_0 = s_0) = P(Y_{t+1} = j | Y_t = i). \tag{3.11}$$

Hence, given the present Y_t and the past $Y_0, ..., Y_{t-1}$ of the process, the future Y_{t+1} only depends on the present and not on the past. The (one-step) transition probabilities from state i to state j are conditional probabilities and defined as

$$p_{ij} = P(Y_{t+1} = j | Y_t = i). \tag{3.12}$$

In a time-homogeneous Markov chain, the one-step transition probabilities are time independent. For a Markov chain with m states, the one-step transition matrix is a stochastic matrix given by

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0m} \\ p_{10} & p_{11} & \cdots & p_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i0} & p_{i1} & \cdots & p_{im} \\ p_{m0} & p_{m1} & \cdots & p_{mm} \end{pmatrix} \tag{3.13}$$

satisfying $p_{ij} \geq 0,$ for all i, j and $\sum\limits_{j \in \mathcal{S}} p_{ij} = 1$.

A Markov chain arises in mortgage terminations since in each consecutive period, the borrower can decide to continue with the contractual payment scheme, prepay the mortgage or default on the mortgage. Full prepayment and default constitute end states, whereas contractual payment can be followed by either of the three states. Figure 2 visualized this Markov chain consisting of $m = 3$ states in a transition diagram.

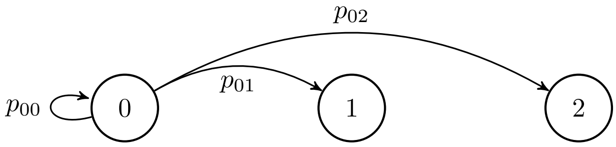


Figure 2: Transition diagram for contractual payment ($Y_t = 0$), full prepayment ($Y_t = 1$) and default ($Y_t = 2$).

The corresponding transition matrix is

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.14}$$

since both full prepayment and default constitute end states.

By including partial prepayments and delinquency status of the borrower, the Markov model can be extended to five states, visualized in Figure 3.

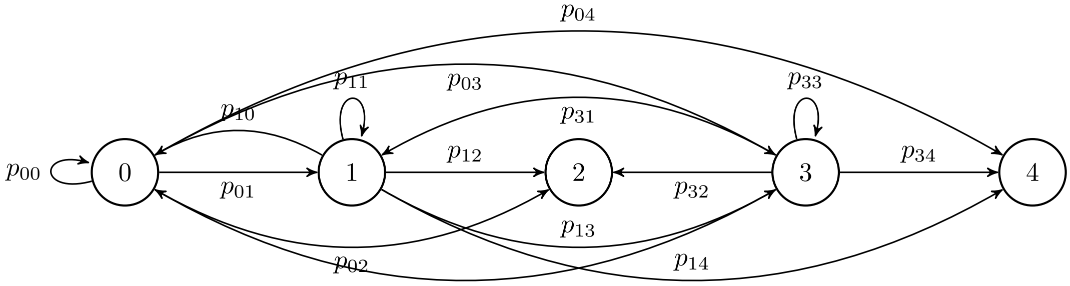


Figure 3: Transition diagram for contractual payment ($Y_t = 0$), partial prepayment ($Y_t = 1$), full prepayment ($Y_t = 2$), delinquency ($Y_t = 3$) and default ($Y_t = 4$).

The corresponding transition matrix is

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} \\ 0 & 0 & 1 & 0 & 0 \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.15}$$

in which full prepayment and default again denote end states. Theoretically, all given transitions in **P** are possible. From a practical perspective, the probability of a transition from partial prepayment to delinquent, p_{13}, will generally be small.

The transition probabilities can be estimated from the data by deriving the likelihood and equating the first order conditions to zero. A consistent MLE estimate for p_{ij} is given by

$$\hat{p}_{ij} = \frac{N_{ij}}{\sum\limits_{j=1}^{m} N_{ij}}, \tag{3.16}$$

in which N_{ij} denotes the number of transitions from i to j.

The transition probabilities can also be estimated using covariate information. By doing this, the model does not exhibit the memoryless property anymore and is therefore more often referred to as a conditional logit model. Estimation can be done by conditioning on a certain (transient) state and performing a logistic regression. A complete estimation of the five state Markov model thus entails estimating a conditional logit model for partial prepayments, delinquent and contractual payment.

3.2 Variable selection procedures

To extract the variables that are most important in determining prepayments, variable selection procedures can be applied. One possibility is a general-to-specific variable selection procedure. Sequentially, variables that are most insignificant are removed from the model after which the logistic regression is run again. This procedure can be conducted such that all variables are significant for one state, all states or for any state. Since the purpose of the present research is to develop a prepayment model a variable selection procedure is applied with which only variables that are significant fr prepayments remain in the model.

Individual significance is assessed by applying a standard t-test. The significance of the categorical variables is assessed jointly, by means of a Wald test. Under the null hypothesis it holds that the set of coefficient estimates, $\hat{\theta}$ is significantly different from zero. The Wald test statistic is given by

$$\frac{\hat{\theta}^2}{var(\hat{\theta})} \sim \chi^2(m), \tag{3.17}$$

in which m denotes the number of elements in $\hat{\theta}$ and the statistic is chi-squared distributed. The significance level for selecting variables is set to ten percent to avoid that too many variables are discarded from the model.

A disadvantage of the general-to-specific approach is that this approach can fail to identify useful predictors in the general model if their significance is affected by irrelevant variables. To prevent important predictors from not ending up in the final model, variables that are relevant based on economic theory are added to the final (specific) to see if the selection is justified.

3.3 Tests for structural breaks

Important diagnostic tests when using models that rely on time series data are tests for parameter stability. Aside from the purpose of assessing model adequacy, such tests are also relevant in providing information on out-of-sample forecasting accuracy. If model parameters are time dependent or exhibit (multiple) structural breaks, forecasting can become challenging. Multiple tests for parameter stability exist. The Chow breakpoint test can be used if there is a known breakpoint in the data set. The disadvantage of this test is that it requires an a priory split of the data. If the location of the break point is uncertain, this test on its own is not very informative. Common tests that are used to determine the timing of the break point are the CUSUM tests and Quandt's likelihood ratio (QLR) test. In this thesis the CUSUM test is used to test for parameter stability since it requires less computation time than the QLR test. Given the sizeable data set this is an important consideration.

3.3.1 CUSUM and CUSUMSQ Tests

The CUSUM and CUSUMSQ test are often used to test for the presence of a structural break (Tanizaki, 2007). The former is useful for detecting the timing of the structural break however the null hypothesis of no structural breaks is too often accepted. Therefore, the CUSUMSQ test is often used to determine the timing of the structural break.

The recursive residual is defined as

$$w_t = \frac{Y_t - x_t\hat{\beta}_{t-1}}{\sqrt{1 + x_t(X'_{t-1}X_{t-1})^{-1}x'_t}}, \tag{3.18}$$

where $X_{t-1} = (x_1, x_2, ..., x_{t-1})$.

The CUSUM test statistic for testing structural change is given by

$$W_t = \sum_{i=k+1}^{t} w_i/\hat{\sigma}, t = k+1, ...T, \tag{3.19}$$

where $\hat{\sigma}^2 = \frac{1}{T-k} \sum\limits_{i=k+1}^{t} w_i^2$ is an unbiased estimate of σ^2. If the disturbance in the regression model is symmetric it holds that $E(W_t) = 0$ and that the distribution of W_t is symmetric around zero. Since the distribution of W_t cannot be obtained explicitly, the test is conducted differently from standard statistical tests. The null hypothesis is accepted if W_t lies within the upper and lower bounds that pass through the points $(k, +/-c_w\sqrt{T-k})$ and $(T, +/-3c_w\sqrt{T-k})$. The parameter c_w depends on the significance level of the test, α. For a significance level of $\alpha = 0.10$, $c_w = 0.850$ (Tanizaki, 2007).

The CUSUMSQ test statistic is defined as

$$S_t = \frac{\sum\limits_{i=k+1}^{t} w_i^2}{\sum\limits_{i=k+1}^{T} w_i^2} \tag{3.20}$$

and the corresponding confidence interval is given by a pair of straight lines $c_s + /-\frac{t-k}{T-k}$ where c_s depends on both the sample size $T-k$ and the significance level α.

If the CUSUMSQ test statistic crosses the boundaries this is an indication of parameter instability. The point where the exceedance takes place, say $t = \tau$, indicates the presence of the structural break. If this is the case, the model can be estimated with the inclusion of an indicator variable such that model parameters are allowed to differ prior and post the structural break. Formally, the test is performed by estimating the logistic regression

$$\ln\left(\frac{P_{itj}}{P_{itk}}\right) = \beta^1_{itj}{}'X_{it}I[t \leq \tau] + \beta^2_{itj}{}'X_{it}I[t > \tau] + \epsilon_{itj}, \tag{3.21}$$

in which β^1_{itj} and β^2_{itj} are coefficients obtained from

$$\ln\left(\frac{P_{itj}}{P_{itk}}\right) = \begin{cases} \beta^1_{itj}{}'X_{it} + \epsilon_{itj}, t = 1, ..., \tau, \\ \beta^2_{itj}{}'X_{it} + \epsilon_{itj}, t = \tau, ..., T. \end{cases} \tag{3.22}$$

The structural break can be included in the model by estimating different coefficient before and after the break date.

3.4 Option theoretic model

Since refinancing of a mortgage is the most important determinant of mortgage prepayment, a separate section will be devoted to the calculation of the refinancing incentive. This section discusses an endogenous prepayment model that determines the optimal refinancing incentive. The refinancing incentive is modelled as if the borrower is able to look back on the evolution of the mortgage rates in the market and determine at which point it would, theoretically, have been optimal to refinance the mortgage. Refinancing is attractive if the market mortgage rate is below the contractual mortgage rate. Figure 4 shows the evolution of the the 30 year fixed rate mortgage (FRM) and the six month EURIBOR over the period 1999M01 to 2014M03.

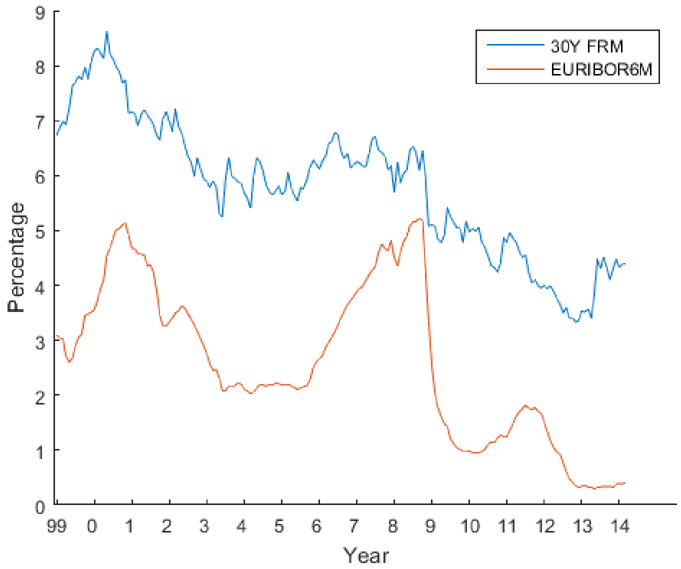


Figure 4: 30Y Fixed Rate Mortgage and EURIBOR6M (1999M01-2014M03).

For a borrower with a contractual mortgage rate of 6.00 percent, looking back from the end of the sample period, it would have been optimal to exercise the option in November 2011 since the difference between the contractual mortgage rate and and the mortgage rate in the market (30Y FRM) is the largest at this point.

The option to refinance can be seen as a European lookback put with a fixed strike price and a varying asset price. The discounted payoff at maturity is given by

$$LBP_{i,T} = P(0,T)E^Q[\max(mo_i - \min_{t \leq \tau \leq T}(ma_\tau), 0)], \tag{3.23}$$

where the strike price is equal to mo_i and is mortgage specific. The put option is is written on ma_t. The payoff of the put option thus depends on two stochastic processes, namely the term structure of the risk free rate and the 30 year market mortgage rate.

Equation (3.23) however fails to incorporate two important features of the optimal refinancing incentive. For one, refinancing in an earlier phase of the mortgage contract is more attractive than refinancing at the same rate close to maturity of the mortgage since more interest payments can be saved. Secondly, mortgage rate differ depending on the remaining maturity of the mortgage which is being refinanced. The contractual mortgage rate has to be compared to the mortgage rate corresponding to the remainder of the maturity at the point of refinancing. The term structure of mortgage rates therefore has to be derived to include the fact that mortgage rates are higher for longer maturities. The outer maximum is the maximum of the put option. The inner maximum refers to the point in time in which the difference between the contractual mortgage rate and the market mortgage rate is the highest.

$$LBP_{i,T} = P(0,T)E^Q[\max(\max[(mo_i - R^{ma}(\tau,T)) * (T - /\tau)], 0)], \tag{3.24}$$

in which $R^{ma}(\tau,T)$ denotes the point on the yield curve where ma_t attains its minimum, that is $ma_t = \min_{t \leq \tau \leq T}(ma_\tau)$. The price of (3.24) can only be obtained numerically.

Both the risk free rate and the market mortgage rate have to be simulated under the risk neutral measure. The two rates can be simulated by separate stochastic processes, assuming a certain correlation between these processes. Given the fact that the evolution of the rates is similar, a straightforward choice would be to simulate the short rate and the mortgage rate with a similar stochastic process. The two rates show the same pattern since the mortgage securities sold by Freddie Mac are implicitly backed by the US government. Therefore, mortgage securities receive an AAA credit rating which in turn puts them in direct competition with risk free rates.

The following section provides a background on dynamic interest rate models.

3.4.1 Interest rate models

The relation between interest rates and the bond price process $P(t,T)$ can be derived from the time t value of $P(T,T) = 1$ as

$$P(t,T) = E^Q[e^{-\int_t^T r(s)ds} | \mathcal{F}_s], \tag{3.25}$$

in which $r(t)$ denotes the short rate and $\mathcal{F}_s$ is the filtration up to $t = s$. Bond prices relate to the term structure, via

$$f(t,T) = -\frac{\ln P(t,T)}{T-t}, \tag{3.26}$$

in which $f(t,T)$ denotes the instantaneous forward rate at time t with maturity T.

Let the dynamics of the short rate be given by

$$dr(t) = \mu(t, r(t))dt + \sigma(t, r(t))d\bar{W}_r(t) \tag{3.27}$$

in which $\mu(t, r(t))$ represents the drift term, $\sigma(t, r(t))$ denotes the diffusion term and $d\bar{W}_r(t)$ is a P Brownian motion with the usual properties (Bjork, 2009). To obtain risk-neural prices for the derivatives, instead of specifying μ and λ under the objective probability measure P, the dynamics of the short rate will be specified under the martingale measure Q. Using the change of measure proposed by Cameron-Martin-Girsanov (CMG) Equation (3.27) can be written as

$$dr(t) = \mu(t, r(t))dt + \sigma(t, r(t))dW_r(t), \tag{3.28}$$

in which $dW_r(t)$ is a Q Brownian motion. Consequently, the family of bond price processes will be determined by the general term structure equation

$$\begin{cases} F_t + (\mu - \lambda\sigma)F_r + \frac{1}{2}\sigma^2 F_{rr} - rF = 0, \\ F(T, r) = 1, \end{cases} \tag{3.29}$$

where λ is defined as the market price of risk. If the term structure $P(t, T)$ has the form

$$P(t, T) = \exp\left(A(t, T) - B(t, T)r(t)\right), \tag{3.30}$$

where $A(t, T)$ and $B(t, T)$ are deterministic functions, then the model is said to possess an affine term structure (ATSM).

Through different specifications of the short rate different term structures can be estimated, which follow directly from Equation (3.29). In the literature, a wide variety short rate specifications can be found. One of the earliest models is the Vasicek model (1977), followed by the Dothan model (1978), the Cox-Ingersoll-Ross model (1985) and the Ho-Lee model (1986). Both the Vasicek model and the CIR model are mean-reverting. The former allows for negative interest rates, while the latter does not. The Dothan model also allows for interest rates to become negative and is not mean reverting. Finally, the Ho-Lee model has a time dependent mean function and allows for negative interest rates.

Desirable properties for the present research are: (1) mean reversion in interest rates and (2) possibility of negative interest rates. Previous research often considers this possibility as a drawback but since in the contemporary economic environments interest rates have become negative, the short rate model has to allow for this possibility. The Vasicek model satisfies these two properties and will be discussed in more depth.

3.4.1.1 Vasicek model

The Vasicek model is a mean reverting process under the risk neutral measure and specifies the dynamics of the short rate using constant coefficients

$$dr(t) = (b - ar)dt + \sigma dW_r(t), \tag{3.31}$$

with $a > 0$ and a mean reversion level of b/a.

The Vasicek term structure is an ATSM with

$$\begin{aligned} B(t, T) &= \frac{1 - \exp\left(-a(T - t)\right)}{a}, \\ A(t, T) &= \frac{(B(t, T) - T + t)(ab - \frac{1}{2}\sigma^2)}{a^2} - \frac{\sigma^2 B^2(t, T)}{4a} \end{aligned} \tag{3.32}$$

.

Due to the fact that the Vasicek model assumes an endogenous term structure, it cannot take the current term structure of interest rates as input (Baldvinsdottir and Palmborg, 2011). Consequently, it is not possible to obtain an exact fit of this model to the market term structure. Since the aim of the ESG is to provide a market consistent valuation, it is desirable that the interest rate model is able to take the current term structure of interest rates as input.

3.4.1.2 Hull-White extension of Vasicek model

The Hull-White extension of the Vasicek model, also called the Hull-White one-factor model (HW1f), allows for an exact fit to the current term structure. The dynamics of the short rate are specified by

$$dr(t) = (\theta(t) - a(t)r(t))dt + \sigma(t)dW_r(t), \tag{3.33}$$

in which the rate of mean reversion $a(t)$ and the diffusion term $\sigma(t)$ are time dependent. Through the term $\theta(t)$ this model can be fitted to the current term structure. The time variation in $\sigma(t)$ allows for an exact fit to the spot or forward volatilities (Plomp, 2013). However, Brigo and Mercurio (2010) (Brigo, 2010) note that if an exact fit to the current term structure is desired it can be dangerous to perfectly fit the volatility term structure as well. The main reason is that the volatility quotes of less liquid markets may be unreliable. To prevent overfitting, the preferred specification of the short rate is obtained by setting $a(t) = a$ and $\sigma(t) = \sigma$ and is given by

$$dr(t) = (\theta(t) - ar(t))dt + \sigma dW_r(t). \tag{3.34}$$

The SDE in (3.34) can be solved by applying Ito's lemma. After observing that the process $r(r)$, conditional on the filtration up to $t = s$, denoted by F_s, the term structure implied by the HW1f model can be derived. The structure of the zero-coupon bond prices is given in (3.25). The term $B(t,T)$ is provided in 3.32 and the term $A(t,T)$ under HW1f is given by

$$A(t,T) = \frac{P(0,T)}{P(0,t)} \exp(B(t,T)f(0,t) - \frac{\sigma^2}{4a} B^2(t,T). \tag{3.35}$$

3.4.1.3 Swaption price under HW1f

The standard Black-Scholes formula for option pricing assumes a constant short rate. In a similar manner, options can be priced using a stochastic short rate, such as the rate implied by the HW1f model. The purpose of this section is to price an option on a bond. The price of a put option at time t with strike K and maturity T on a zero coupon bond with maturity S is given by

$$ZBP(t,T,S) = KP(t,T)\Phi(-h + \sigma_P) - P(t,S)\Phi(-h), \tag{3.36}$$

where

$$\begin{aligned} \sigma_P &= \sigma \sqrt{\frac{1 - e^{2a(T-t)}}{2a}} B(t,S), \\ h &= \frac{1}{\sigma_P} \ln \frac{P(t,S)}{P(t,T)K} + \frac{\sigma_P}{2}. \end{aligned} \tag{3.37}$$

A European payer swaption with strike price K gives the holder the right at maturity T to enter into a payer interest rate swap. A payer interest rate swap allows the owner to exchange a fixed rate for a floating rate at a number of future dates, $T_1, T_2, ..., T_n$. The time between these dates, δ, is generally fixed. The maturity date of the swap usually coincides with the first reset date of the underlying swap, $T = T_0$. A European swaption can be viewed as a portfolio of put options on a pure discount bond. The price of a payer swaption with maturity T is given by

$$PS(t, T, T_i, N, K) = N \sum_{i=1}^{n} K\delta ZBP(t, T, T_i, K_i), \tag{3.38}$$

in which $K_i = A(t, T_i)e^{-B(t,T_i)r*}$ where $r*$ is the value of the spot rate at time T for which

$$\sum_{i=1}^{n} K\delta A(t, T_i)e^{-B(t,T_i)r*} = 1. \tag{3.39}$$

3.4.1.4 Model calibration

In order to simulate interest rate paths, values for a and σ in (3.34) have to be determined. This can be done by optimizing the value for $\hat{a}$ and $\hat{\sigma}$ such that the HW1f model is best fitted to market prices. The best fit is defined as the minimization of the sum of the squared relative deviation between market prices for swaptions and swaption prices implied by the HW1f model, given in (3.39). The objective function is given by the minimum of

$$O = \sum_{i} \sum_{j} [PS(t, T_j, T_i, N, K) - BPS(t, T_j, T_i, N)]^2 \tag{3.40}$$

where i and j are different tenors and maturities, respectively.

To determine the swaption prices implied by the market, the model is calibrated to market data. Calibration of the HW1f model requires two sets of prices, namely (i) the current term structure to which $\theta(t)$ can be fitted and (ii) the short rate process $r(t)$. For the former, discount factors of zero coupon bonds can be used. In the literature at the money (ATM) swaption prices are generally used to determine the discount factors due to the fact that these instruments are widely traded.

In the market, swaptions are quoted in terms of their volatilities. To determine the swaption prices from these quotes Black's formula can be used. Black's formula for the price of a payer swaption is given by (Plomp, 2013)

$$BPS(t, T_i, N) = N\delta(F(t)\Phi(d_1(t)) - K\Phi(d_2(t))) \sum_{i=1}^{n} P(t, T_i), \tag{3.41}$$

where

$$\begin{aligned} d_1(t) &= \frac{\ln\left(\frac{F(t)}{K}\right) + \frac{1}{2}\sigma(t)^2(T_0 - t)}{\sigma(t)\sqrt{T_0 - t}}, \\ d_2(t) &= d_1(t) - \sigma(t)\sqrt{T_0 - t}. \end{aligned} \tag{3.42}$$

Here $\Phi(.)$ denotes the standard normal distribution, $F(t)$ represents the future swap rate, $\sigma(t)$ is the quoted swaption volatility and $i = 1, ...n$ denote the future reset dates of the swap. A derivation of the Black formula can be found in Bjork (2009) (Bjork, 2009).

3.5 Assessing model performance

Since the aim of this thesis is to determine the model that is most appropriate for predicting prepayment rates, measures have to be identified that are capable of assessing model performance. After the initial models are estimated, a variable selection procedure is conducted and diagnostic tests are applied, model performance measures can be computed.

3.5.1 Panel data performance

Model performance measures can be applied after estimating the logistic regression

$$\ln\left(\frac{P_{itj}}{P_{itk}}\right) = X'_{it}\beta_{itj} + \epsilon_{itj} \tag{3.43}$$

in which $P_{itj} = P(Y_{it} = j)$ is the probability of event j occurring and $P_{itj} = P(Y_{it} = k)$ is the probability of the benchmark event k occurring.

Measures of fit for logistic regressions generally fall into two categories, namely measures that look at predictive power and goodness of fit tests (Allison, 2014). Examples of the former are measures for R^2. However, no consensus is reached on the best manner that this can be calculated for a logistic regression. The McFadden R^2 or the pseudo R^2, is a likelihood ratio test that determines the proportional reduction in error variance of an intercept only model versus a model that includes explanatory variables. The Cox and Snell R^2 is a more general version of the former, however it has an upper bound less than one (Allison, 2014). Another example is the Tjur R^2 which has the advantage of being closely related to linear models and has an upper bound of one. However, it does not depend on the likelihood. Therefore it could be the case its value decreases after the addition of explanatory variables. Concluding, the McFadden R^2 appears to be the most appropriate R^2 for a logistic regression

$$R^2 = 1 - \ln(L_M)/\ln(L_0), \tag{3.44}$$

in which L_M denotes the likelihood of the model including regressors and L_0 is the likelihood of the model with only an intercept. The likelihood is given by

$$L = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{m} Y_{itj} * \ln(\hat{P}_{it} = j). \tag{3.45}$$

For L_M, $(\hat{P}_{it} = j)$ is given in Equation (3.2) and for L_0, $(\hat{P}_{it} = j) = \sum\limits_{i=1}^{n}\sum\limits_{t=1}^{T} Y_{itj}/N$, in which N is the total number of observations.

Examples of goodness of fit measures are the Pearson goodness of fit measure and the deviance measure (Allison, 2014) of which the latter is the most widely used, given by

$$D = 2\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{m} Y_{itj} \ln\left(\frac{Y_{itj}}{n_{it} * (\hat{P}_{it} = j)}\right). \tag{3.46}$$

Another approach to determine the relative quality of different model specifications is by means of information criteria such as the Akaike Information Criterion (AIC) and the Schwartz Information Criterion (SIC). These criteria are well suited for finding a balance between model fit and model parsimony. They distinguish useful risk drivers from irrelevant ones these criteria penalize the estimation of additional parameters. The AIC is defined as

$$AIC = 2k - 2\ln L, \tag{3.47}$$

in which k denotes the number of parameters in the model and L is the likelihood. The SIC is given by

$$SIC(k) = T\ln\hat{\sigma}^2 + k\ln T \tag{3.48}$$

and penalized the inclusion of additional regressors more than the AIC.

To facilitate a comparison between the three- and five state models, the goodness of fit measures are based on the contribution only of the three states (contractual payment, full prepayment and delinquent) to the likelihood. To this end, the number of observations used in the computation of the likelihood in the five state models has been adjusted.

3.5.2 Cross sectional analysis

Aside from looking at model performance in a panel data set, it can also be assessed over time t or over mortgages i. Cross sectional model performance is assessed by transforming the panel dataset into a cross-sectional dataset by averaging information per mortgage over time. The logistic regression in Equation (3.43) looks as follows

$$\begin{aligned} &\sum_{t=1}^{T_i} \ln\left(\frac{P_{itj}}{P_{itk}}\right)/T_i = \sum_{t=1}^{T_i} X'_{it}\beta_{itj}/T_i + \sum_{t=1}^{T_i} \epsilon_{itj}/T_i \\ &\ln\left(\frac{P_{ij}}{P_{ik}}\right) = X'_i\beta_{ij} + \epsilon_{ij} \end{aligned} \tag{3.49}$$

in which T_i denotes the number of (time series) observations per mortgage i. This number is mortgage specific since the panel is unbalanced.

Since in a cross sectional out-of-sample analysis only one estimate can be made per individual i, a contingency table is more appropriate to assess model performance. A contingency table compares realized versus predicted state realizations. Predicted realizations are obtained by comparing the estimates probability for mortgage i being in state j at time t, $\hat{P}_{itj}$, to the in-sample probability for state j. The in-sample probability is given by

$$P_j = N(Y_j)/N \tag{3.50}$$

in which $N(Y_j)$ denotes the number of observations classified as state j and N denotes the total number of observations. The predicted value of the dependent variable reads $\hat{Y}_{it} = j$ if the following holds

$$\hat{P}_{itj} > P_j \tag{3.51}$$

for states $j = 1, ..., m$. It could be the case that in-sample thresholds for multiple states are breached. To ensure only one state assignment for an observation, the state can be assigned which has the highest relative breach severity

$$\frac{\hat{P}_{itj} - P_j}{P_j}. \tag{3.52}$$

Table 2 shows a contingency table of observed states, Y_j, and forecasted states, $\hat{Y}_j$ for state $j = 1, ...m$.

Y_{itj} $\hat{Y}_{itj}$	$j = 1$	$j = 2$	...	$j = m$	
$j = 1$	$n(\hat{Y}_1 Y_1)$	$n(\hat{Y}_1 Y_2)$	...	$n(\hat{Y}_1 Y_m)$	$N(\hat{Y}_1)$
$j = 2$	$n(\hat{Y}_2 Y_1)$	$n(\hat{Y}_2 Y_2)$	...	$n(\hat{Y}_2 Y_m)$	$N(\hat{Y}_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$j = m$	$n(\hat{Y}_m Y_1)$	$n(\hat{Y}_m Y_2)$	...	$n(\hat{Y}_m Y_m)$	$N(\hat{Y}_m)$
	$N(Y_1)$	$N(Y_2)$	...	$N(Y_m)$	N

Table 2: Contingency table of predicted and realized state classifications.

The success rate for state j is defined as

$$S_j = \frac{n(\hat{Y}_j Y_j)}{N(Y_j)}, \tag{3.53}$$

the false alarm rate is defined as the number of times the model predicts state j when the actual state is k

$$F_{jk} = \frac{n(\hat{Y}_j Y_{k \neq j})}{N(Y_{k \neq j})}. \tag{3.54}$$

Finally, the missed rate is defined as the number of times the model predicts state k when in fact the state is j

$$M_{jk} = \frac{n(\hat{Y}_{k \neq j} Y_j)}{N(Y_j)}. \tag{3.55}$$

Cross sectional model performance can be assessed both in-sample and out-of-sample. A cross-sectional out-of-sample analysis can be conducted by using covariate information from all mortgages except the i'th mortgage and using the coefficient estimates to form a prediction on the state of mortgage i.

3.5.3 Time series analysis

For a time series analysis the panel data set can be transformed into a time series data set by averaging different mortgages over calender months in Equation (3.43) as

$$\begin{aligned} \sum_{i=1}^{n_t} \ln\left(\frac{P_{itj}}{P_{itk}}\right)/n_t &= \sum_{i=1}^{n_t} X_{it}'\beta_{itj}/n_t + \sum_{i=1}^{n_t} \epsilon_{itj}/n_t \\ \ln\left(\frac{P_{tj}}{P_{tk}}\right) &= X_t'\beta_{tj} + \epsilon_{tj} \end{aligned} \tag{3.56}$$

in which n_t denotes the number of mortgages at time t. This number varies per month since each month new mortgages can enter via an origination or can leave if they are matured, prepaid or defaulted on.

For the present research a contingency table is only used to assess cross sectional model performance as opposed to for evaluating time series performance. Since there are no restrictions on the amount of end states a model can predict (e.g. prepayments and defaults) the model can predict multiple end states for a certain mortgage while in the dataset a mortgage is no longer observed if an end state occurs. Therefore, using contingency tables for assessing time series model performance provides a too negative view of the actual model performance. The models often signal a prepayment a few months before the actual prepayment occurs. This feature cannot be captured in the contingency tables. Time series model performance can be assessed based on (i) unbiasedness; (ii) accuracy and (iii) efficiency. The unbiasedness of model forecasts is evaluated by means of a simple hypothesis test of a zero mean in the prediction errors. The prediction errors in a logistic regression are given by

$$\epsilon_{tj} = P_{tj} - \hat{P}_{tj}, \tag{3.57}$$

in which P_{tj} is the in-sample probability for state j and the residuals follow an Extreme Value distribution, $\epsilon_{tj} \sim EV$. When the sample size is large, the Central Limit Theorem (CLT) can be used to obtain the asymptotic distribution of the residuals. If the model is well specified, the distribution of the residuals converges to a normal distribution for large sample sizes, e.g. $\epsilon_{tj} \sim NID(0, \sigma_j)$.

To test whether this assumption is valid, Kolmogorov-Smirnov (KS) test for normality can be used. The test statistic, $\mathcal{D}_j$, is defined as

$$\mathcal{D}_j = sup_{\epsilon_{tj}} |F_n(\epsilon_{tj}) - F_0(\epsilon_{tj})| \tag{3.58}$$

in which F_n is the empirical CDF of the residuals and F_0 is the CDF of a normal distribution with the same mean and variance. If this test does not reject that the residuals are asymptotically normally distributed, a simple t-test can be used to test whether the prediction error has a zero mean.

Another desirable property of any time series model is that there is no time variation in the residuals. To test whether this is the case, residuals can be regressed on their lagged value(s) as

$$\epsilon^*_{tj} = \rho_1 \epsilon^*_{t-1,j} + \rho_2 \epsilon^*_{t-2,j} u_t j + ... + \rho_l \epsilon^*_{t-l,j} \tag{3.59}$$

in which l denotes the lag in empirical autocorrelation. To determine the number of lags by which possible autocorrelation can be captured, the Partial Autocorrelation Function (PACF) of the residuals can be plotted. The PACF (not the ACF) can be used for this purpose since it corrects the time series for autocorrelation at lower lags. By means of the Ljung-Box (LB) test it can be formally tested whether the residuals exhibit there autocorrelation. The LB test is specified as

$$LB = T(T+2) \sum_{l=1}^{L} \frac{\hat{\rho}_l^2}{T-l} \sim \ \chi^2(L) \tag{3.60}$$

in which T denotes the number of time series observations.

The accuracy of model predictions can be determined by computing the Mean Squared Predictor Error (MSPE) as

$$MSPE_j(T) = \frac{1}{T} \sum_{t=1}^{T} (P_{tj} - \hat{P}_{tj})^2. \tag{3.61}$$

The model with the lowest MSPE is preferred.

The efficiency of model predictions can be determined though the Mincer-Zarnovic regression. Ideally, it should not be possible to predict the prediction error itself with any information available at time t. This implies that in a regression of the error term on the fitted value

$$\epsilon_{tj} = \alpha_0 + \alpha_1 \hat{P}_{tj} + \eta_{tj} \tag{3.62}$$

α_0 should be close to zero and α_1 should be close to one.

Time series analysis can be conducted both in-sample and out-of-sample. Out-of-sample time series forecasts can be made using a Moving Window (MW) or an Expanding Window (EW). In the former approach the size of the window remains the same whereas with the window size grows for forecasts further away in the future. A MW approach is more appropriate if the parameters show time variation. The time series out-of-sample performance measures are similar to the in-sample measures. The forecast errors for forecast horizon h are given by

$$\epsilon_{T+h|T} = P_{T+h} - \hat{P}_{T+h|T}, \tag{3.63}$$

where the subscript j is omitted here but the forecast errors are state specific.

4 Data description

The model is estimated using a panel data set of mortgage loans provided by Freddie Mac, the Single Family Loan-Level data set. It contains loans that originated between January 1, 1999 and September 31, 2013. The data set consists of monthly observations on fixed rate fully amortizing mortgages with a maturity of thirty years. The sample is geographically dispersed in the United states and covers loans that were originated in 51 different states. The fixed rate interest rate period coincides with the full maturity of the mortgage contract. Unlike common in the Netherlands, the data set does not contain intermediate periods in which the interest rate is reset. The data set includes 50,000 loans that originated each year between 1999 and 2013. The total number of mortgage loans in the data set amounts 737,111. Each loan is tracked from origination date until mortgage termination or maturity with as cutoff date March 2014. Followed over time this leads to an unbalanced panel data set consisting of 31,018,317 observations.

Since computation time is too lengthy when all observations are included in the model, a random sample is drawn. A set of 10,000 mortgages is sampled uniformly at random without replacement from the 737,111 mortgages. The mortgages are tracked over time which leads to a total of 409,319 observations.

4.1 Dependent variables

The dependent variable used in the models is a categorical variable indicating the state of the principal payment scheme. There are two classifications for the dependent variable, namely one for the three state models and one for the five state models.

4.1.1 Three state models

The three states denote contractual payment of the mortgage loan, full prepayment of the mortgage and default. The first category is used as a benchmark category. The benchmark category, $Y_{it}^3 = 1$, denotes contractual payment of the remaining outstanding principal. The category $Y_{it}^3 = 2$ denotes a voluntary prepayment of the full remaining outstanding principal. This classification is derived from the data set by means of a variable indicating the reason for a zero balance code. The category $Y_{it}^3 = 3$ denotes default. Loans with a delinquency status of more than 90 days are classified as default. Observations on mortgages after a delinquency status of more than 90 days are removed from the data set. This leads to a total of 29,932,667 observations in the full data set. The default category also comprises foreclosures by an alternative group, for example though a short sale, third party sale, charge off or note sale. In this case the borrower is unable to make principal or interest payments and the property can be seized. A repurchase prior to a property disposition and a Real-Estate-Owned (REO) disposition are also classified as default. A REO disposition occurs if the lender becomes the owner of the property after an unsuccessful foreclosure auction.

Table 3 provides the number of observations for the dependent variable per category for both the full data set and the random sample of 10,000 mortgages.

Y_{it}^3	**Description**	**Full data set**	**Random sample**
1	Contractual payment UPB	29,400,423	402,154
2	Voluntary prepayment of full UPB	490,828	6,616
3	Default	41,416	549

Table 3: Description of dependent variable in three state models.

The percentage of observations classified as prepayments in the sample denotes 1.616 percent, compared to 1.640 percent in the full data set. For default the figures are 0.134 percent and 0.138 percent, respectively. Since these numbers are comparable the sample can be used as an approximation for the entire data set.

Table 3 can also be interpreted on loan level. From the 10,000 mortgages 6,616 mortgages are prepaid at some point whereas 549 default at some point in the time period January 1999 until March 2014. The high prepayment rate underwrites the importance of appropriately modelling the propensity to prepay.

The Markov models are obtained by determining transition probabilities from state sequences for each loan. Both full prepayment and default constitute end states, whereas the other categories are transient states. If an end state of a loan is observed, then the state sequence for that loan is observed entirely. However, if a loan is not fully prepaid nor defaulted within the sample period the remaining states in the lifetime of the loan are not observed. Since the data set is right-censored, the number of observations for each transition is less than the total number of observations. Table 4 shows the number of observations for each possible transition for the three-state Markov model.

$Y^3_{i,t+1}$ Y^3_{it}	**1**	**2**	**3**
1	392,161	6,609	540
	(0.9821)	(0.0166)	(0.0014)
2	0	0	0
	(0)	(1)	(0)
3	0	0	0
	(0)	(0)	(1)

Table 4: Number of observations (percentages) per transition in three-state Markov model.

The number of observations for transitions starting form an end state is zero since the mortgage contract is terminated after the end state is observed. The number of transitions to a certain state is less than the number of observations for a certain state if the state is the first observation for a mortgage. For example for prepayments, there are seven mortgages for which the first observation is classified as a prepayment.

4.1.2 Five state models

The five state dependent variable consists of additional categories for partial prepayments and delinquency status. The delinquency state is derived from the number of days the borrower is delinquent, based on the due date of the last payment installment (DDLPI) reported by servicers to Freddie Mac. The delinquency state varies between zero and 119 days. An observation is classified as delinquent if the borrower is delinquent between 30 and 90 days on his mortgage payments. If a borrower is delinquent from 1 to 29 days, this observation is classified as contractual payment.

Partial prepayments are defined as prepayments in excess of the payments expected under the contractual payment scheme, often also referred to as curtailments. The actual cash flows can be identified from the data set by means of the current actual Unpaid Principal Balance (UPB) which is given for each observation. This reflects the mortgage ending balance as reported by the servicer for the corresponding monthly reporting period. The UPB for the first six months after the loan origination is censored. To obtain observations for this initial period, the UPB for these months is interpolated from the original loan balance and the first given observation on the actual UPB. Hence, it is assumed that in this initial period no curtailments take place.

The monthly contractual cash flows for fully amortizing loans are generated as follows

$$CF = \frac{P * r}{1 - (1 + r)^n}, \tag{4.1}$$

in which r denotes the monthly mortgage rate and n is the original term of the loan. The mortgage rate is constant over the lifetime of the mortgage. For each month, the difference between the actual and contractual payment is calculated. Next, for each loan, the average excess payment and the standard deviation thereof is computed over the lifetime of the loan. Consequently, an observation is classified as a curtailment if the actual payment exceeds the contractual payment by three times its standard deviation.

Table 5 provides the number of observations for the dependent variable per category for both the full data set and the random sample of 10,000 mortgages. The number of full prepayments and default are identical to those reported for the three state models. The percentage partial prepayments denotes 2.59 percent in the random sample and 2.57 percent for the full data set. For delinquent payments the figures are 1.32 percent and 1.43 percent, respectively. Again, the percentages are comparable.

Y^5_{it}	**Description**	**Data set**	**Random sample**
1	Contractual payment UPB	28,207,383	386,239
2	Unscheduled partial return of UPB	769,904	10,596
3	Voluntary prepayment of full UPB	484,119	6,514
4	Delayed payment of less than 90 days	429,845	5,421
5	Default	41,416	549

Table 5: Description of dependent variable in five state models.

Table 6 shows the number of observations for each possible transition for the five-state Markov model.

$Y^5_{i,t+1}$ Y^5_{it}	**1**	**2**	**3**	**4**	**5**
1	366,119	9,592	6,228	3,129	537
	(0.9495)	(0.0249)	(0.0162)	(0.0081)	(0.0014)
2	7,390	686	239	103	0
	(0.8779)	(0.0815)	(0.0284)	(0.0122)	(0.000)
3	0	0	0	0	0
	(0)	(0)	(1)	(0)	(0)
4	2,737	318	40	2,189	12
	(0.5168)	(0.0600)	(0.0076)	(0.4133)	(0.0023)
5	0	0	0	0	0
	(0)	(0)	(0)	(0)	(1)

Table 6: Number of observations (percentages) per transition in five-state Markov model.

In Table 6 it can be seen that the number of transitions from partial prepayments to default, $P(Y^5_{t+1} = 5 | Y^5_t = 2)$, is zero. This means that the five state Markov model conditioned on partial prepayments is in fact a four state model. Furthermore, the table shows that the number of observations for some transitions is limited. The transition probability $P(Y^5_{t+1} = 4 | Y^5_t = 2)$ for example is extremely small. Intuitively this makes sense since if a person is delinquent on its mortgage payments in the current period, full repayment of the remaining outstanding principle in month later is indeed unlikely. Furthermore, partial prepayments mostly are an indicator of more partial prepayments in the next period and sometimes even for full prepayment.

4.1.3 Time series analysis

Figure 5 plots the dependent variable over the period January 1999 until March 2014 (in months). The dashed line (green) indicates the number of observations in each calender month. This number increases over time since more and more mortgages are purchased by Freddie Mac. At the end of the sample period the number of observations decreases again due to the fact that mortgages have either expired (after 30 years), have been prepaid or have defaulted. The rates in the dependent variable are constructed by dividing the number of observations on a certain category for a certain calender months, $N_t(O_j)$, by the total number of observations in that month, N_t.

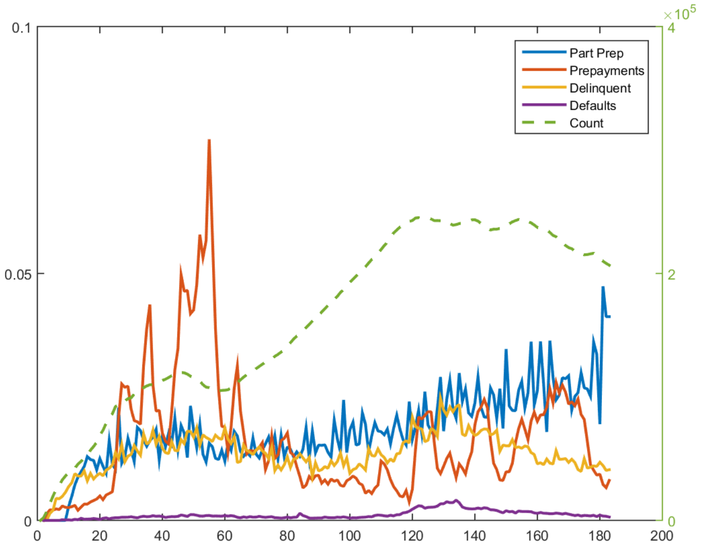


Figure 5: Dependent variable over time (percentages), 1999M01-2014M03.

Prepayment rates (red) appear to show persistence over time as periods of high and low prepayment rates can be distinguished. In the period June 2001 until December 2003 prepayment rates are the highest after which they decrease to a minimum level in 2008. This decrease can be linked to a deterioration of the US economy during the credit crisis of 2008. After this decrease prepayment rates fluctuate around a lower level. In comparison to prepayments, default rates (blue) are lower over the entire sample. Again the crisis is visible by the increase in default rates shortly after 2008. In the five state models also partial prepayments (blue) and delinquency status (yellow) are incorporated. Partial prepayments are more fluctuating around a slightly increasing trend over the sample period and do not seem to follow the same pattern as full prepayments. The number of mortgages with a delinquency status seems to be on the same trend as defaulted mortgages, albeit in a more exaggerated manner. Two peaks are visible of which the highest again during 2008.

4.2 Independent variables

Table 7 provides a description of explanatory variables for mortgage i and time t that are included in the Freddie Mac data set. Complementary, four macroeconomic risk drivers have been added. State level information is used if available, otherwise region specific data are used[4]. The unemployment rate is available on a monthly basis and is added to the data set on state level. The mortgage rate in the market is also provided on a monthly basis and is refined by US region. Divorce rates and US housing prices are included on a monthly basis per state. The categorical variables Loan Age, Property Type, Loan Purpose and Region are included as dummy variables.

[4] States are classified into five regions according to the Federal Reserves' classification: North Central, North East, South East, South West and West.

Variable	**Description**
loanage_{it}	The number of months since the note origination month of the mortgage.
mortrate_{it}	The current interest rate on the mortgage note, taking into account any loan modifications.
FICO_i	A credit score, between 301 and 850, that indicates the creditworthiness of the borrower and is indicative of the likelyhood that the borrower will timely repay future obligations.
firsthome_i	Indicates whether the property is the first home bought by the borrower.
insurance_i	The percentage of loss coverage on the loan (between 0 and 55 percent).
DTI_i	Debt to income (DTI) ratio defined as the sum of the borrower's monthly debt payments divided by the tot monthly income used to underwrite the borrower at the origination date of the mortgage.
loansize_i	The original UPB scaled by the mean original UPB of the selected sample.
LTV_i	The original Loan-To-Value (LTV) ratio defined by the original prinicpal divided by the purchase price of the mortgaged property.
penalty_i	Dummy variable that indicates whether the borrower is obliged to pay a penalty for unscheduled return of the principal.
occupancyowner_i	Variable indicating that the mortgage is owner occupied, investment property or a second home.
propertytype_i	Denotes whether the property is a condominium (CO), planned unit development (PUD), cooperative share (CS), manufactured home (MH) or single family home (SFH).
purpose_i	Indicates whether the mortgage is a cash-out refinance mortgage, no cash-out refinance mortgage or a purchase mortgage.
state_i	Indicates the state within which the property securing the mortgage is located.
unempl_t	Seasonally adjusted unemployment level per US state.
mortgagerate_t	Mortgage rate on 30-year fully amortizing fixed rate mortgages available in the market per US region[5].
divorcerate_t	Divorce rates per US state.
housingprice_t	Seasonally adjusted indexed home price levels per US state.

Table 7: Description of explanatory variables.

Table 8 shows summary statistics for the risk drivers for the random sample of 10,000 mortgages. Summary statistics for the full sample of 737,111 mortgages can be found in Appendix B.

Variable	Mean	St. Dev.	Min	Max
Loan Age	32.529	28.728	1	181
FICO score	730.482	53.930	530	832
Firsthome	0.138	0.345	0	1
Mortgage insurance	5.045	10.445	0	40
DTI	34.166	11.740	1	65
Loansize	0.915	0.512	0.080	4.187
LTV	72.877	15.931	6	100
Penalty	0.001	0.029	0	1
Unemployment	6.143	2.089	2.100	14.900
Divorcerate	4.051	0.877	1.700	9.900
Houseprice	155.655	26.342	100.590	206.670
Refinancing	-0.612	0.985	-6.020	2.870
Property Type Condo	0.001	0.701	0	1
Property Type Planned Unit Development	0.277	0.447	0	1
Property Type Cooperative Share	0.302	0.459	0	1
Property Type Manufactured Housing	0.243	0.429	0	1
Property Type Single Family Home	0.180	0.384	0	1
Purpose Cash-Out	0.649	0.477	0	1
Purpose No Cash-out Refinance	0.094	0.292	0	1
Purpose Purchase	0.257	0.437	0	1
Region 1	0.070	0.255	0	1
Region 2	0.001	0.033	0	1
Region 3	0.142	0.349	0	1
Region 4	0.008	0.089	0	1
Region 5	0.779	0.415	0	1
Loan Age below 1Y	0.293	0.455	0	1
Loan Age 2-3Y	0.212	0.409	0	1
Loan Age 3-4Y	0.150	0.357	0	1
Loan Age 4-6Y	0.183	0.386	0	1
Loan Age 6-8Y	0.094	0.291	0	1
Loan Age 8-10Y	0.046	0.209	0	1
Loan Age 10-15Y	0.022	0.148	0	1

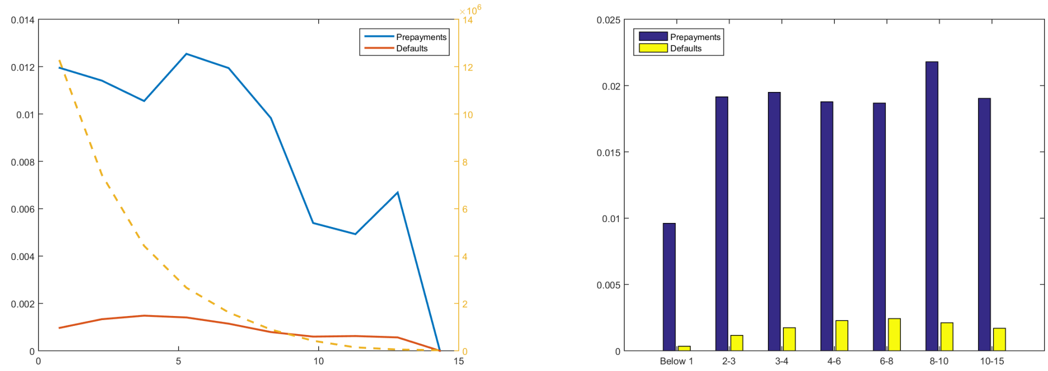
Table 8: Summary statistics of dependent variables for sample of 10,000 mortgages (409,319 observations).

4.3 Bivariate analyses

In this section, bivariate analyses are conducted to determine (non)linear effects of the different risk drivers on the dependent variable, most notably prepayment rates. In this way, it can be assessed whether risk drivers can directly be included in the models or should be included with a transformation. The figures for the continuous variables are constructed by combining observations in bins. The bin size determines the degree of smoothing visible in the figures. Decreasing the number of bins can aid in revealing a trend. The number of bins used, k, is provided below each figure. The explanatory variables will be discussed in the same order as in Table 7.

4.3.1 Loan Age

Loan age is tracked on a monthly basis and varies between 1 and 181 months. Figure 6(a) displays the seasoning of the mortgages for the full sample for prepayments (blue) and defaults (red). Prepayments are at the highest level for loan ages of about six years after which they sharply decrease to peak again for loan ages of twelve years. Default rates increase steadily during the first few years after loan origination. After a small decrease at ages of 10 years default rates rise again.



(a) Loan Age Continuous (k=10)

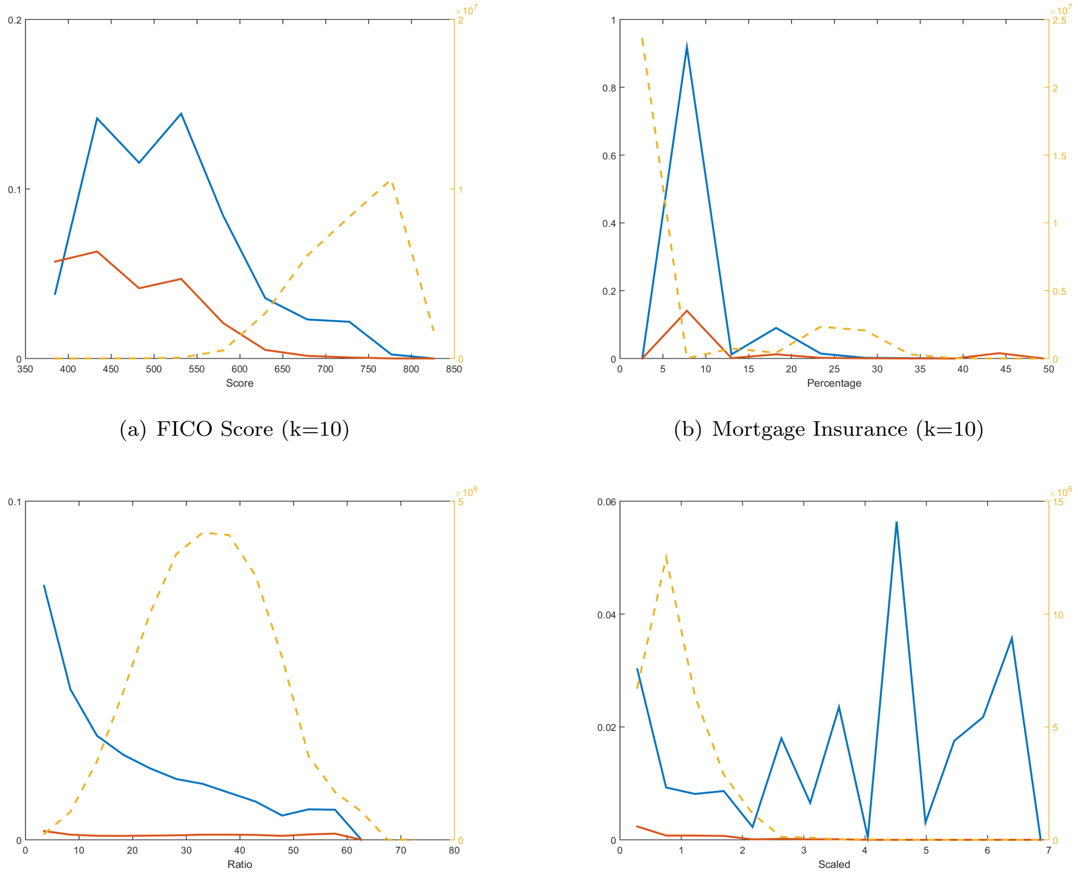
(b) Loan Age Categorical

Figure 6: Bivariate analyses of loan age (in years) on prepayment rates and default rates for the full data set.

Since the effect of loan age on prepayment rates is not linear over the lifetime of the mortgage, this variable is included as a categorical variable by buckets. The bucket size is be determined by ensuring the presence of a similar amount of observations in each bucket. Buckets of smaller size are statistically less reliable. Since there a more observations for younger loans, visible by the dashed yellow line in Figure 6(a), these buckets are narrower. Figure 6(b) shows prepayment and default rates per bucket.

4.3.2 Loan specific risk drivers

Figures 7(a) through 7(d) show the bivariate analyses for the variables FICO score, Mortgage Insurance, DTI and Loan Size on prepayment (blue) and default rates (red).



(a) FICO Score (k=10)

(b) Mortgage Insurance (k=10)

(c) Debt-to-Income (k=15)

(d) Loan Size (k=15)

Figure 7: Bivariate analyses of loan specific risk drivers on prepayment rates (blue) and default rates (red) for the full data set.

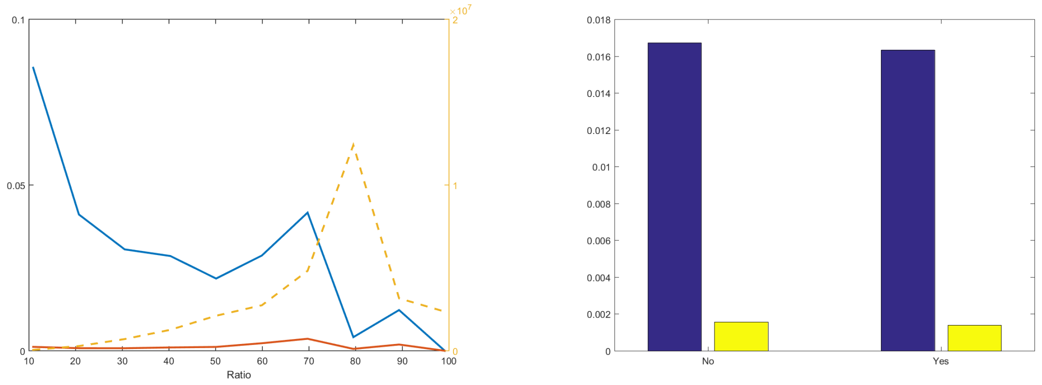
A higher FICO score is associated with a higher degree of creditworthiness. Consequently, it is expected that this score is inversely related to the default rate. This is confirmed in Figure 7(a). The relation between the FICO score and prepayment rates also seems to be negative for FICO scores exceeding 550. Prepayment rates are relatively high for borrowers with a low creditworthiness. Bearing in mind the limited number of observations for FICO scores below 550 (indicated by the dashed line) the FICO score is included in the model with no transformations.

The theoretical relationship between mortgage insurance and default rates is negative. The higher the percentage of the mortgage that is covered, the lower the probability that the mortgagee will default. This is visible in Figure 7(b). Both defaults and prepayments peak at mortgage insurance rates of about eight percent. Given the limited number of observations for this point, it could be the case that only a few observations cause this spike. Hence, again this variable will be included in the model without transformations.

A higher DTI ratio is associated with a higher risk involved in the ability to meet all scheduled payments. Therefore, its relation with the default rate is expected to be positive. This is confirmed in Figure 7(c). A higher DTI ratio also generally entails that people have fewer free resources to make unscheduled excess payments. Hence, its relation with prepayment rates is expected to be negative. In Figure 7(c) this is clearly visible by the decreasing trend in the blue line. Since this effect is more or less linear for the majortity of the observations the DTI ratio is included with no transformations.

Loan size is a scaled variable indicating the dollar size of the loan relative to the average loan size in the sample. The effect of loan size on prepayment and default rates is less evident (Figure 7(c)) although it appears that prepayment rates are higher for larger loans. For mortgages with a below average size prepayment rates are slightly higher, indicated by the decreasing line before loan size is equal to one. This is as expected since the amount of funds required for a full prepayment of the UPB is relatively lower.

Figures 8(a) and 8(b) show the bivariate analyses of the loan-to-value ratio (LTV) and the first home indicator.



(a) Loan-to-Value (k=10)

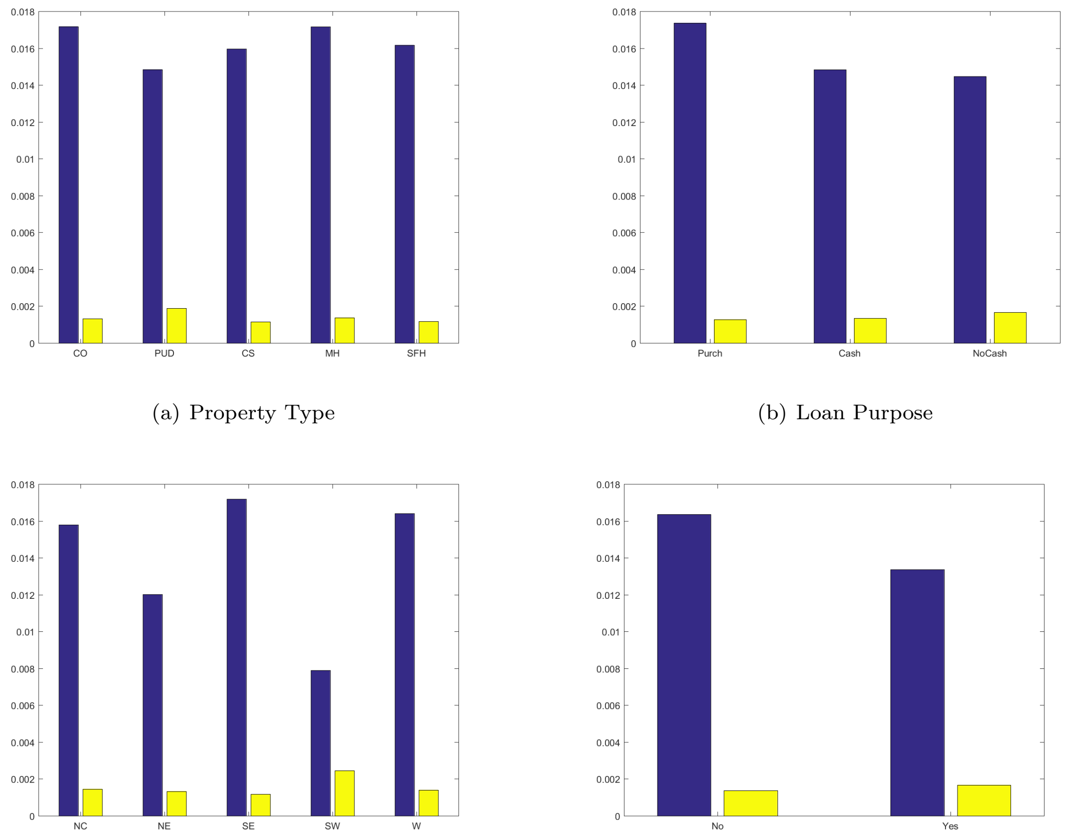
(b) First Home

Figure 8: Bivariate analyses of loan specific risk drivers on prepayment rates (blue) and default rates (red/yellow) for the full data set.

A higher LTV is indicative of a higher default risk since mortgagees have the possibility to walk away from the loan if the value of the residential property is significantly lower than the outstanding loan. This is visible in Figure 8(a) by the slightly increasing default rates for higher LTVs. The relation between LTVs and prepayment rates is expected to be negative since a lower value of the residential property is associated with lower wealth of the mortgagee, especially given the fact that a residential property constitutes the largest fraction of wealth for an individual. This effect is visible in the figure by the on average decreasing trend for prepayments.

It is expected that mortgagees for which the property is a first home have a higher default rate due to its relation with job insecurity and age of the borrower. The effect of a first home on prepayment rates is twofold. On the one had it is expected that prepayment rates are lower since young mortgagees often do not have a lot of spare funds to finance a prepayment. On the other hand, young people often relocate more than older people which could lead to higher prepayment rates.

Figures 9(a) through 9(d) show the bivariate analyses for the occupancy status of the residential property, the purpose of the mortgage, the region in which the residential property is located an whether a prepayment penalty applies to the mortgage contract on prepayment rates (blue) and default rates (yellow).



(a) Property Type

(b) Loan Purpose

(c) Region

(d) Prepayment Penalty

Figure 9: Bivariate analyses of loan specific risk drivers on prepayment rates (blue) and default rates (yellow) for the full data set.

In Figure 9(a) it is visible that prepayment rates are slightly higher when the property is a condominium (CO) or a manufactured home (MH) whereas default rates are quite a bit higher when the property is a planned unit development (PUD) or a single family home (SFH).

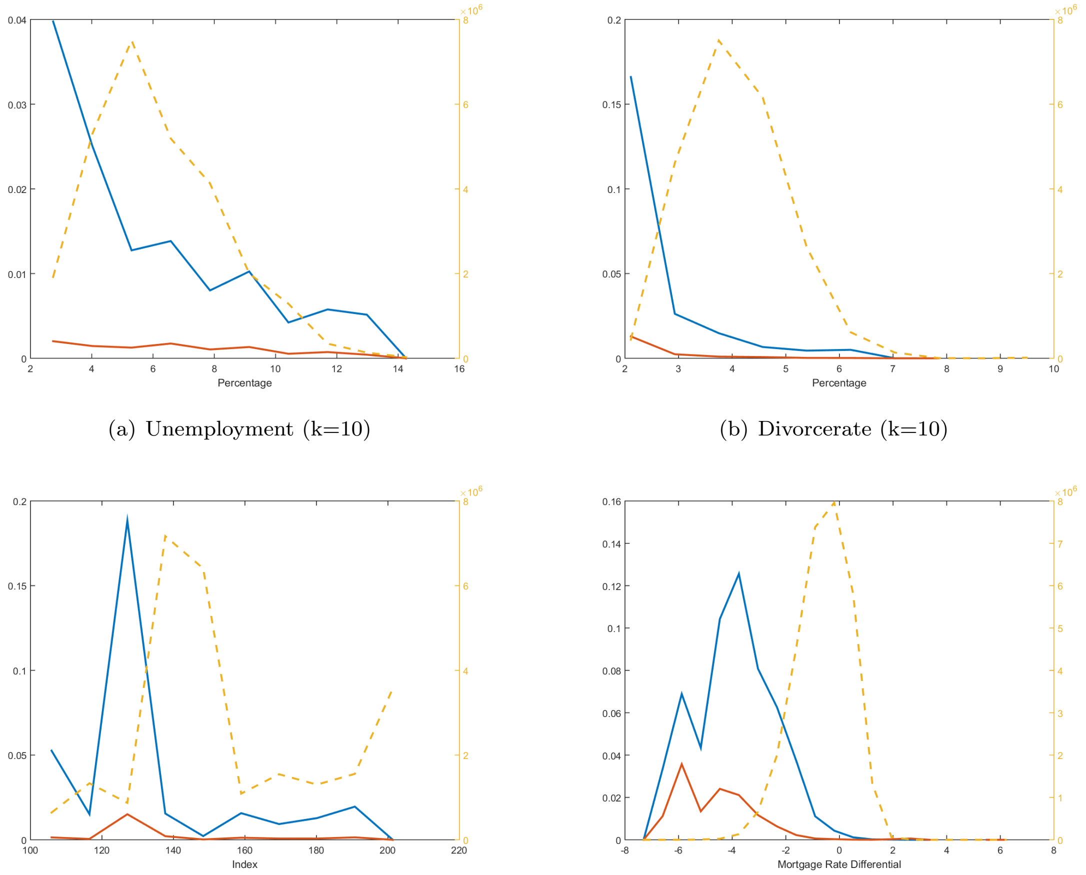
When the purpose of the loan is is to purchase the property, prepayment rates are slightly higher compared to when the loan purpose is to cash out (Figure 9(b)). Default rates are considerably higher when the purpose of the loan is to not cash out.

Figure 9(c) shows that prepayment and default rates differ per region. In region 4, the South West, default rates are higher and prepayment rates are lower compared to the rest of the US. Prepayment rates are the highest in the South East (region 3).

Figure 9(d) shows the difference in prepayment and default rates when a prepayment penalty applies to the mortgage and when it does not. In the presence of a prepayment penalty it is expected that prepayment rates are lower, which is confirmed in the figure. The presence of a prepayment penalty does not influence the default rate.

4.3.3 Macroeconomic variables

Figures 10(a) through 10(d) show the bivariate analyses for the included macroeconomic variables unemployment rate, divorce rate, house price index and market mortgage rates on prepayment (blue) and default rates (red).



(a) Unemployment (k=10)

(b) Divorcerate (k=10)

(c) House Price (k=10)

(d) Refinancing Incentive (k=20)

Figure 10: Bivariate analyses of macroeconomic variables on prepayment rates (red) and default rates (blue) for the full data set.

The relation between unemployment rates and default rates is expected to be positive, whereas a negative relation with prepayment rates is expected. In Figure 10(a) especially the negative relation between unemployment levels and prepayments is visible.

The theoretical relation between the divorce rate and the prepayment rate is positive. A divorce often leads to relocation which in turn can be a trigger to pay off the current mortgage before maturity. In Figure 10(b) this relationship is not confirmed.

The house price index is constructed as the house price index relative to the starting month, January 1999 in which the index is 100. If house prices increase, it is expected that borrowers relocate more frequently. In Figure 10(c) is can be seen that on average, prepayment rates are lower when house prices have increased with the excpetion of the sharp peak in prepayments (and defaults) at house prices which have increased by 25 basis points.

The refinancing incentive is one of the most important determinants for prepayment. Refinancing is optimal if the mortgage rate currently paid on the mortgage is lower than the rate available in the market. Figure 10(d) shows the difference between the market mortgage rate and the contractual mortgage rate. If this difference is negative, prepayment is optimal which is clearly visible in the figure. Moreover, it appears that the refinancing incentive does not influence the prepayment rate in a linear manner. Taking into account the fact that the gain from refinancing is higher when the remaining UPB is higher, the refinancing incentive is included in the model by also incorporating information on the UPB.

The refinancing incentive is derived as the difference in discounted UPB payments at the market rate and the contractual mortgage rate. If the present value of the remaining principal payments discounted at the current mortgage rate is higher than the present value of the remaining principal payments discounted with the best available rate in the market, then refinancing at the latter rate is optimal. The refinancing incentive is given by

$$RFI_{it} = \frac{PV_{it}^{ma} - PV_{it}^{mo}}{PV_{it}^{ma}}, \tag{4.2}$$

in which PV_{it}^{mo} denotes the present value of the remaining principal payments plus interest payments at the mortgage contract rate, given by

$$PV_{it}^{mo} = \sum_{t=1}^{T-k} \frac{P_{it} * mo}{(1+r_t)^{T-t}}. \tag{4.3}$$

The current mortgage rate is assumed to be constant over the contractual life of the mortgage. The term PV_{it}^{ma} in Equation (4.2) denotes the present value of the remaining principal payments plus interest payments at the mortgage market rate, given by

$$PV_{it}^{ma} = \sum_{t=1}^{T-k} \frac{P_{it} * ma_t}{(1+r_t)^{T-t}}, \tag{4.4}$$

in which the available market mortgage rate differs per month.

5 Results

5.1 Multinomial logit model

5.1.1 Three state MNL model

The MNL model is initially estimated using the entire set of explanatory variables, see Appendix C Table 35. To extract the variables that are most important in determining prepayments a general-to-specific variable selection procedure for prepayments is applied. Individual significance as well as joint significance of the categorical variables is assessed at a significance level of 10 percent. Table 9 shows the coefficient estimates and p-values of the MNL3 model.

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
C	-14,457	0,000	-19,166	0,000
Firsthome	-0,444	0,005	-0,339	0,110
Mortgage insurance	-0,019	0,001	-0,013	0,064
Loansize	-0,165	0,062	-0,043	0,727
LTV	0,015	0,000	0,052	0,000
Houseprice	0,022	0,000	0,024	0,000
Refinancing	-1,414	0,000	-1,633	0,000
Loan Age		0,000		0,000
Property Type		0,000		0,000
Loan Purpose		0,000		0,000
Region		0,000		0,000

Table 9: Coefficient estimates and p-values MNL3 model.

The variables Property Type, Loan Purpose, Region and Loan Age are included as categorical variables. Table 9 reports the p-values of the Wald test. The majortity of these variables is individually insignificant but jointly constitute an important indicator for prepayments. Individual coefficients estimates and p-values are given in Appendix E Table 43. This table also includes information on dummies that were added during the estimation porcedure to accont for missing values. Rows with missing values are skipped entirely in the estimation procedure of the MNL coefficients. To avoid this, dummy variables are added that have the value of one whenever the observation for an explanatory variable is missing. Consequently, the missing values of the explanatory variables are set to zero.

In Table 9 it can be seen that all variables are significant for predicting prepayments at the ten percent level. Variables that are removed in the variable selection procedure are FICO score, DTI ratio, Prepayment Penalty, Unemployment rate and Divorce rate.

5.1.2 Five state MNL model

The five state MNL model (MNL5) includes the transient states partial prepayment and delinquent payment aside form the end states (full) prepayment and default. Contractual payment is again chosen as reference category. Table 10 shows the results of the MNL5 model after applying the general-to-specific variable selection procedure. The general model is given in Appendix C Table 36. The variable FICO score, the First Home indicator, the DTI ratio, Loan Size and Prepayment Penalty have been removed during this procedure. Appendix E Table 44 shows the individual coefficient estimates an p-values of the categorical variables as well as for the included missing value dummies.

	Part.Prep		**Prep.**		**Delinq.**		**Default**	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-5,520	0,000	-17,022	0,000	-7,232	0,000	-22,215	0,000
Mortgage insurance	0,004	0,004	-0,015	0,005	0,004	0,015	-0,014	0,055
LTV	0,001	0,512	0,010	0,007	0,016	0,000	0,039	0,000
Unemployment	0,019	0,001	0,174	0,000	-0,010	0,191	0,347	0,000
Divorcerate	0,066	0,000	-0,238	0,021	-0,033	0,040	-0,329	0,004
Houseprice	-0,001	0,000	0,020	0,000	0,004	0,000	0,027	0,000
Refinancing	-0,107	0,000	-1,170	0,000	-0,558	0,000	-1,382	0,000
Property Type		0,000		0,000		0,000		0,081
Loan Purpose		0,000		0,000		0,000		0,000
Region		0,243		0,000		0,000		0,000
Loan Age		0,000		0,000		0,000		0,000

Table 10: Coefficient estimates and p-values MNL5 model.

Although the majority of the included variables is significant for both prepayment and partial prepayment, the variables LTV and Region are insignificant for the latter. This indicates that these two states can not necessarily be estimated with the same set of risk drivers. Comparing Table 10 and 9, some common prepayment risk drivers can be identified, namely the variables Mortgage Insurance, LTV, House Price and Refinancing Incentive. In the MNL5 model two additional macro economic variables are present whereas in the MNL3 model the indicator for First Home and the variable Loan Size are significant.

5.1.3 Independence of Irrelevant Alternatives Property

To validate the IIA property, the Hausman-McFadden test explained in Section 3.1.1 is performed. This is done by removing defaults as a category from the dependent variable and deleting observations from the data set that are classified with this state. Consequently, the logistic regression is performed on the binary variable in case of the three state MNL model and a multinomial logistic regression in case of the five state MNL model. The coefficient estimates are provided in Appendix D Tables 41 and 42. The HM test probabilities for both MNL models for all states under a $\sim\ \chi^2(36)$ are almost equal to zero. Therefore, it is concluded that the IIA property is validated in both the MNL3 model and the MNL5 model.

5.1.4 Cross sectional analysis

To investigate the performance of the model over over different mortgages, a cross sectional analysis is conducted. To this end covariate information of mortgages is averaged over time periods, as explained in Equation (3.49). The predicted states are defined by comparing the predicted probability for mortgage i for state j, $\hat{P}_{ij} = \sum_{t=1}^{T_i} \hat{P}_{itj}/T_i$ to the in-sample probability for state j, P_j. See Equations (3.50)-(3.52).

Tables 11 and 12 show the contingency tables for the MNL3 and MNL5 models based on a cross sectional regression of 10,000 mortgages. Rows indicate predicted states and columns indicate observed states. The diagonal represents the number of correctly predicted states.

$\hat{Y}_{it}^3$ Y_{it}^3	**1**	**2**	**3**	
1	2750	420	36	3206
2	80	5662	347	6089
3	5	534	166	705
	2835	6616	549	10000

Table 11: Contingency table of predicted and realized state classifications for MNL3 model.

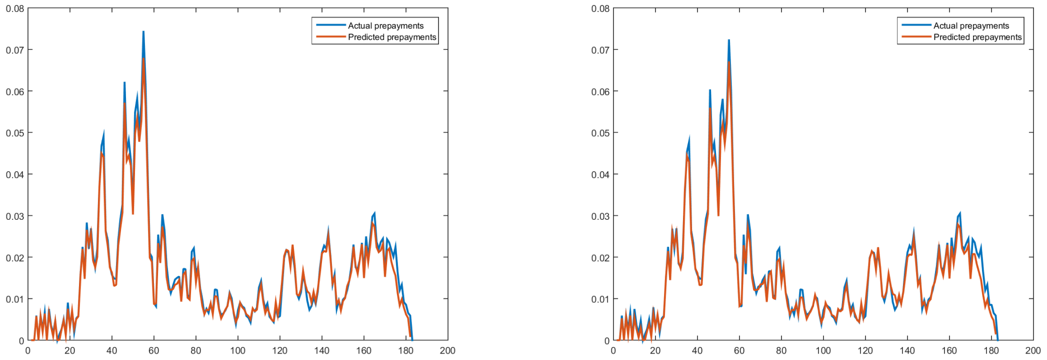
$\hat{Y}_{it}^5$ Y_{it}^5	**1**	**2**	**3**	**4**	**5**	
1	216	65	97	2	4	384
2	381	2010	541	25	49	3006
3	28	84	5175	76	320	5683
4	3	5	14	2	0	24
5	6	14	687	20	176	903
	634	2178	6514	125	549	10000

Table 12: Contingency table of predicted and realized state classifications for MNL5 model.

The tables show the cross sectional performance of the MNL models. The section on model performance, Section 6.1, will investigate the performance in more detail by computing the success rate, false alarm rate and missed prepayments rate based on the contingency tables.

5.1.5 Time series analysis

In Section 4.1.3 it was observed that prepayment rates are time varying. More specifically, periods of high prepayments and low prepayments were visible in Figure 5. This section will investigate to what extent the MNL models are capable of capturing this time varying effect. To this end, the logistic regression for the panel dataset is transformed into a time series dataset according to Equation (3.56). Figures 11(a) and 11(b) show the predicted prepayment probabilities, $\hat{P}_{tj} = \sum_{i=1}^{n_t} \hat{P}_{itj}/n_t$ of the MNL models versus the realized prepayments on a monthly basis for $t = 1999M01, ..., 2014M03$.



(a) MNL3 prepayment estimates

(b) MNL5 prepayment estimates

Figure 11: Time series analysis of MNL prepayment estimates.

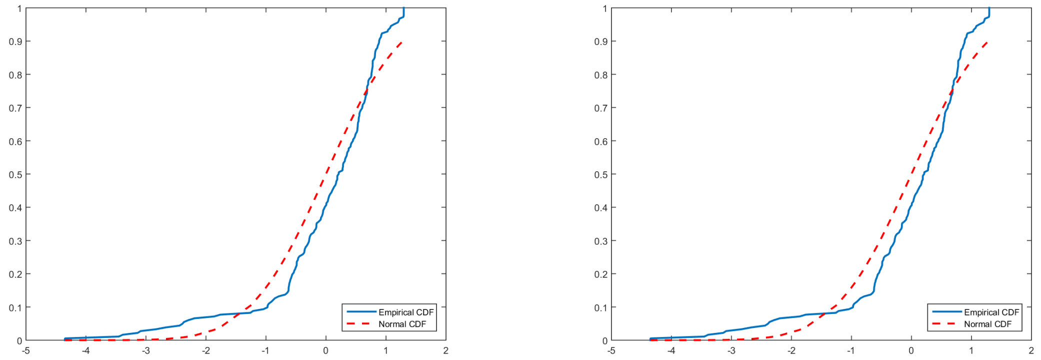
In Figures 11(a) and 11(b) it can be seen that the models mimic the actual prepayment rates over the sample period. Both models slightly underestimate the prepayment rate in the period November 2000 until October 2003 and near the end of the sample period, from March 2013 to March 2014. However, overall the models appear to be well capable of capturing the general pattern in prepayments over time.

Comparing the two figures, it can be seen that they are very similar. Hence, it appears that including partial prepayments as an additional category in the MNL model is not necessary to accurately predict prepayment rates over time.

5.1.6 Residual diagnostics

The residuals for the MNL models are derived according to Equation (3.57). Given the large number of observations in the present data set the CLT is used to infer whether the distribution of the prepayment residuals converges to a normal distriution. The number of mortgage observations per calender month varies between 2,246 in 1999M02 and 260,449 in 2011M10 and can be found in Figure 5.

Figures 12(a) and 12(b) plot the empirical CDF of the prepayment residuals of the MNL models together with a normal CDF with the same mean and variance (given in Table 13).



(a) MNL3 ECDF prepayment residuals

(b) MNL5 ECDF prepayment residuals

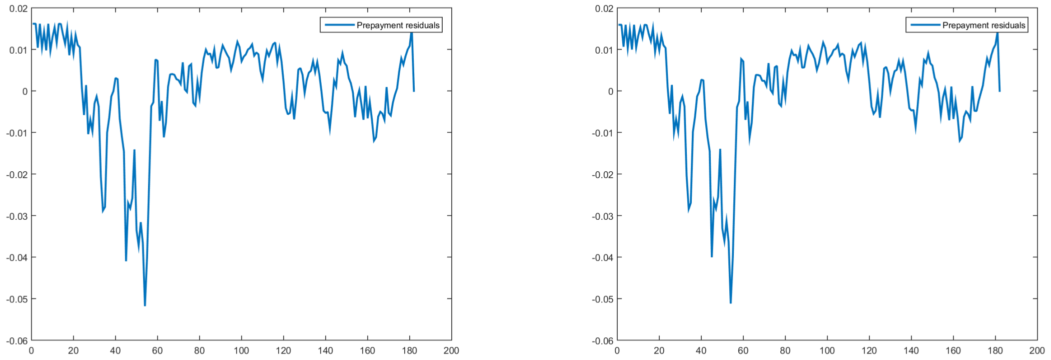
Figure 12: Empirical and normal CDF of prepayment residuals in MNL models.

The figures show that the ECDF resembles a normal CDF but is less smooth. Table 13 provides p-value for the Kolomogorov-Smirnov (KS) test for normality which rejects that the time series residuals are normally distributed. However the time series residuals are based on less observations (182 observations) compared to the panel data residuals (409,3119 observations). The panel data residuals are normally distributed with a mean equal to zero as indicated by the p-value for the t-test in Table 13.

	$\bar{\mu}(\epsilon_{jt}) * 10^3$	$\hat{\sigma}^2(\epsilon_{jt}) * 10^3$	p_{KS}	p_{ttest}
MNL3	0,566	12,070	0,006	0,529
MNL5	0,658	11,933	0,008	0,459

Table 13: Mean, variance, KS statistic and p-value zero mean hypothesis for prepayment residuals MNL models.

Figures 13(a) and 13(b) plot the prepayment residuals, ϵ_{tj} of the MNL models over the sample period.

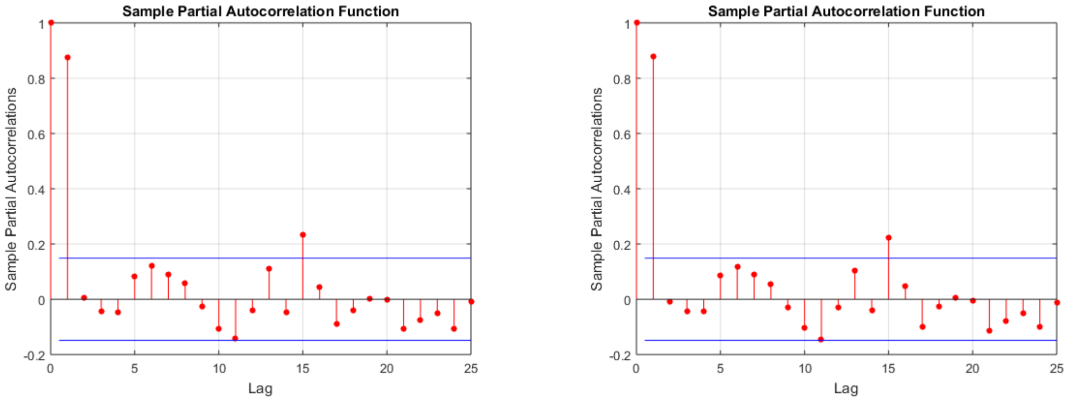


(a) MNL3 prepayment residuals

(b) MNL5 prepayment residuals

Figure 13: Time series analysis of MNL3 and MNL5 prepayment estimates.

The residual plots show the underestimation of the prepayment rates in the period November 2000 until October 2003, the period in which prepayment rates are at the highest level. Furthermore, in Figures 13(a) and 13(b) it can be seen that there is some time variation in the residuals.To determine the order of residual autocorrelation, the Partial Autocorrelation Functions (PACF) are plotted in Figures 14(a) and 14(b). To limit computation time forecasts are made for a sample of 200 mortgages.



(a) MNL3 PACF prepayment residuals

(b) MNL5 PACF prepayment residuals

Figure 14: Partial Autocorrelation Functions MNL prepayment residuals.

The prepayment residuals of both the MNL3 and the MNL5 model exhibit significant first order autocorrelation. This is confirmed by the p-values of the Ljung-Box test on the AR(1) coefficient of the residuals under a $\sim\ \chi^2(1)$ in Table 14.

	$\hat{\rho}_{\epsilon_1}$	$LB(1)$
$MNL3$	0.884	0.000
$MNL5$	0.888	0.000

Table 14: Residual autocorrelation and Ljung-Box test statistic for MNL3 and MNL5 prepayment residuals.

In an attempt to control for autocorrelation in the MNL models, dummy variables for month and/or year have been added. This slightly decreased the AR(1) coefficient however comes at the expense of increased computation time due to the addition of a large number of variables. Another possible remedy that has been examined is the inclusion of the macro-economic variables in first differences. However, this did not lower the autocorrelation and diminished the accuracy of the predictions.

In OLS regressions, residual autocorrelation is often remedied by including a lagged dependent variable as explanatory variable in the model. In the three state MNL model this is not possible directly since a mortgage drops out of the data set after it has been prepaid or defaulted on. A possible solution is to include the percentage of mortgages that is in state j in the preceding month(s) as an endogenous regressor. This has been done for both the MNL3 model and the MNL5 model. The residual autocorrelation decreased slightly (from 0.884 to 0.817 in the MNL3 model) and the model estimates remained relatively similar. One drawback of including an endogenous regressor in the model is that this comes at the expense of the significance of other variables in the model. This was observed by the fact that in the variable selection procedure only a few variables turned out to be significant for prepayments whereas the endogenous regressor was highly significant. Moreover, including an endogenous regressor leads to inconveniences when making multiple period ahead predictions. Since the aim of this thesis is to develop a prepayment model that is capable of predicting prepayments not only in-sample but also out-of-sample (this is of more interest for financial institutions), the endogenous regressors are not included in the final models.

5.1.7 Forecasting

The cross sectional out-of-sample model performance is assessed by transforming the panel dataset into a cross-sectional dataset according to Equation (3.49). The out-of-sample analysis is conducted by using covariate information from all mortgages except the i'th mortgage and using the coefficient estimates to form a prediction on the state of mortgage i. The contingency tables for the MNL3 forecasts and the MNL5 forecasts are provided in Tables 15 and 16 respectively. The results of the cross sectional forecasts will be discussed in more depth in Section 6.2.1.

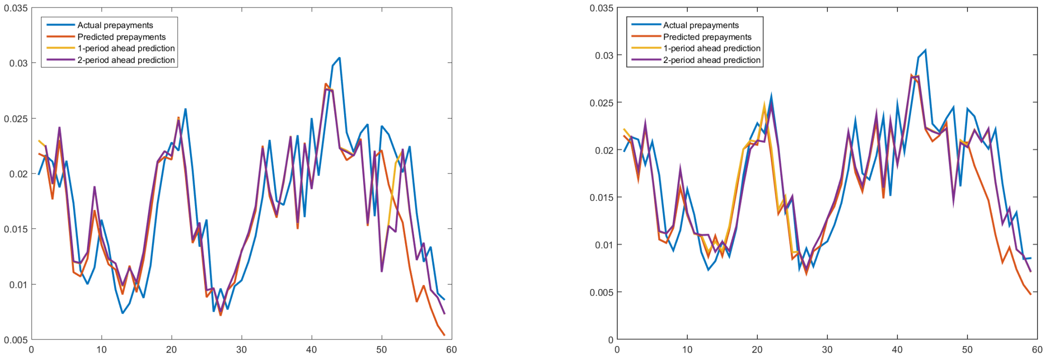
Y_{it}^3 \ $\hat{Y}_{it}^3$	**1**	**2**	**3**	
1	52	14	1	66
2	3	108	6	117
3	0	10	7	17
	54	132	14	200

Table 15: Out-of-sample contingency table for MNL3 model.

Y^5_{it} \ $\hat{Y}^5_{it}$	**1**	**2**	**3**	**4**	**5**	
1	8	0	3	0	0	11
2	10	31	16	0	1	58
3	0	4	102	3	9	118
4	0	0	0	0	0	0
5	0	0	9	0	4	13
	18	35	130	3	14	200

Table 16: Out-of-sample contingency table for MNL5 model.

The time series out-of-sample performance of the models is assessed by means of a one period out of time analysis. For the one period out-of-time analysis the sample is split at two thirds (January 2009) such that the in-sample period contains 122 months and the forecast sample contains 61 months. The one period ahead forecasts are made using an Expanding Window (EW). Figures 15(a) and 15(b) show the forecasted prepayment probabilities along with the actual prepayments and in-sample prepayment predictions for the one-step ahead forecasts of the MNL3 and MNL5 models.



(a) MNL3 prepayment forecasts.

(b) MNL5 prepayment forecasts.

Figure 15: One -and two period ahead prepayment predictions MNL models, January 2009 to March 2014.

In Figure 15(a) it can be seen that the one period ahead prediction of the MNL3 model is relatively accurate and follows the prepayment pattern over the forecast period quite well. One exception is is in March 2013 in which the forecast shows a sharp drop. The one step ahead forecast of the MNL5 model in Figure 15(b) also appears to be accurate. The forecasts of both models seem to overestimate the prepayment rate in the beginning of the forecast period (January 2009) whereas they underestimate prepayments towards the end of the sample.

5.2 Competing risk model

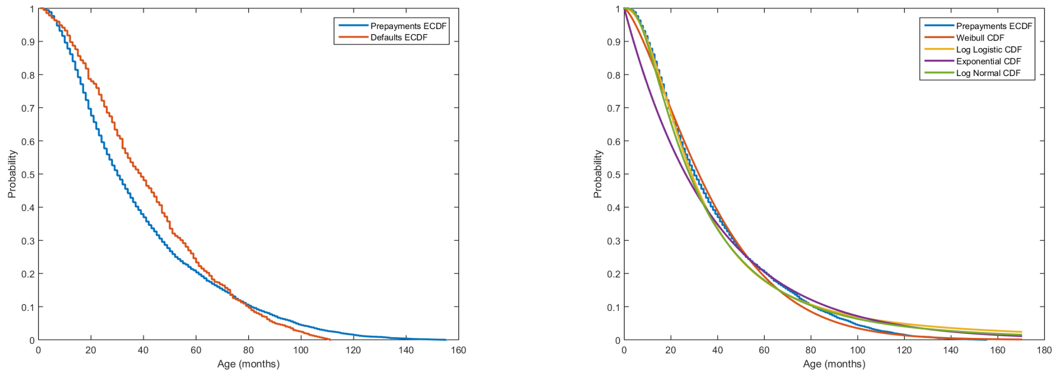
In case there are no covariates, the survivor function in the competing risk model can be estimated with the Kaplan-Meier (K-M) estimator (Rodriguez, 2005). Let $t_{j1} < t_{j2} < ... < t_{jk_j}$ denote the k_j

distinct failure times of type $j = 1, 2$. Let n_{ji} denote the number of loans at risk of being terminated for reason j just before t_{ji} and let d_{ji} denote the number of loan terminations due to cause j at time t_{ji}. Censored data are not included in this analysis.The K-M estimator of the survivor function is given by

$$\hat{S}_j(t) = \coprod_{i:t_{ji}<t} \left(1 - \frac{d_{ji}}{n_{ji}}\right) \tag{5.1}$$

and can be interpreted as the probability that a mortgage is terminated at time t_{ji} given that it was not terminated before this time. The conditional probability of mortgage termination at time t

Ignoring censored cases, the KM estimate coincides with the empirical surival function. The empirical survival functions for mortgage termination due to prepayments and defaults are given in Figure 16(a). The figure shows the survival probability of loans that prepay at some point (blue) or default at some point (red). There are discontinuities at observed times of mortgage termination. Since loans are only tracked in the time period 1999-2014, the data are censored and for the majority of the loans that have not been prepaid or defaulted in this period it is not possible to determine their actual maturity.



(a) ECDFs prepayments and defaults.

(b) ECDF prepayments and theoretical CDFs.

Figure 16: Empirical and theoretical survivor functions for prepayments and defaults.

Figure 16(a) demonstrates that the survival rate for defaults is higher for younger loans compared to the survival rate for prepayments. Figure 16(a) also illustrates that the relation between loan age and the survival probability of the mortgage is non-linear. This effect can be captured by including dummy variables for loan age.

As explained in Section 3.1.2, the baseline hazard rate λ_{j0} can be fitted to various theoretical distributions. Figure 16(b) shows the ECDF for prepayments together with the CDF of the Weibull, Log Logistic, Exponential and Log Normal distribution. The fitted parameters are given in Table 17.

Parameters	Weibull	Log Logistic	Exponential	Log Normal
$\hat{p}_1$	41,641	3,362	37,866	3,330
$\hat{p}_2$	1,382	0,478		0,834

Table 17: Fitted theoretical distributions to prepayment survival rates.

The prepayment survival rate can best be modelled by a Weibull distribution with location parameter $\hat{p}_1 = 41.641$ and scale parameter $\hat{p}_2 = 1.382$. To incorporate non linearities in the relation between loan age and prepayment rates, this variable will be included as a categorical variable.

5.3 Markov models

5.3.1 Three state Markov model

The Markov models are obtained by determining transition probabilities from state sequences for each loan. Both full prepayment and default constitute end states, whereas the other categories are transient states. Including covariate information to estimate the transition probabilities can be done by estimating the MNL conditional on a certain state. Table 18 reports the results after variable selection. Table 37 in Appendix C shows the result of the general model and Appendix E Table 45 provides individual estimates of the categorical variables.

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
C	-5,485	0,000	-4,243	0,000
FICO score	0,001	0,000	-0,011	0,000
Firsthome	-0,106	0,015	0,004	0,978
DTI	-0,006	0,000	0,030	0,000
Loansize	0,465	0,000	0,388	0,000
LTV	-0,003	0,000	0,028	0,000
Unemployment	-0,096	0,000	0,080	0,000
Divorcerate	0,099	0,000	-0,020	0,699
Refinancing	-0,689	0,000	-0,487	0,000
Property Type		0,000		0,002
Loan Purpose		0,018		0,002
Region		0,000		0,000
Loan Age		0,000		0,000

Table 18: Coefficient estimates and p-values Markov3 model.

5.3.2 Five state Markov models

The five state Markov model incorporates information on intermediate states. Coefficients are estimated by conditioning on the transient states and conducting a logistic regression. This is done for contractual payment, partial prepayment and delinquent payment.

Conditional on partial prepayment, the results of the Markov model are given in Table 19. For the Markov models a general-to-specific variable selection procedure is applied by which the specific model only contains variables that are significant for predicting prepayments at the 10 percent level. Table 38 in Appendix C shows the estimation output for the Markov5(1) model before variable selection.

Part.Prep	Part.Prep		Prep.		Delinq.	
	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-2,158	0,000	-16,015	0,818	-4,606	0,000
Loansize	-0,444	0,000	0,323	0,025	-0,468	0,060
Penalty	-13,485	0,985	2,095	0,064	-11,414	0,988
Unemployment	0,013	0,500	-0,099	0,001	0,054	0,228
Refinancing	0,117	0,013	-0,339	0,000	-0,168	0,138
Property Type		0,000		0,000		0,000
Loan Purpose		0,014		0,435		0,430
Region		0,325		0,000		0,000
Loan Age		0,000		0,000		0,000

Table 19: Coefficient estimates and p-values Markov5(1) model.

Table 19 again reports the Wald statistic for the categorical variables. The coefficients and p-values of the individual variables as well as for the included dummies to account for missing values are given in Appendix E Table 46.

Conditional on delinquent payments, the results of the Markov model after variable selection are given in Table 20. The general Markov5(5) model can be found in Appendix C Table 39 and the estimates for the individual categorical variables in Appendix E Table 47.

Delinq.	Part.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-2,439	0,007	-3,229	0,169	-0,837	0,059	-5,596	0,167
FICO score	0,000	0,997	0,006	0,046	-0,002	0,001	-0,001	0,872
LTV	-0,006	0,198	-0,024	0,015	0,006	0,013	0,021	0,480
Unemployment	0,032	0,321	-0,424	0,000	0,038	0,018	0,250	0,088
Refinancing	-0,030	0,660	-0,298	0,091	-0,159	0,000	0,225	0,528
Property Type		0,210		0,865		0,000		0,000
Loan Purpose		0,807		0,108		0,000		0,542
Region		0,007		0,056		0,000		0,120
Loan Age		0,000		0,000		0,000		0,000

Table 20: Coefficient estimates and p-values Markov5(3) model.

Conditional on contractual payment, the results of the Markov model after variable selection are given in Table 21. The general Markov5(5) model can be found in Appendix C Table 38 and the estimates for the individual categorical variables in Appendix E Table 48.

Contr.Pay	Part.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-5,396	0,000	-5,412	0,000	3,512	0,000	-3,956	0,000
FICO score	-0,001	0,000	0,001	0,000	-0,015	0,000	-0,012	0,000
Firsthome	0,057	0,120	-0,087	0,052	-0,099	0,136	0,020	0,902
DTI	0,000	0,671	-0,007	0,000	0,014	0,000	0,030	0,000
Loansize	0,096	0,000	0,476	0,000	-0,017	0,690	0,400	0,000
LTV	0,001	0,099	-0,003	0,000	0,011	0,000	0,029	0,000
Unemployment	0,092	0,000	-0,081	0,000	0,028	0,002	0,068	0,001
Divorcerate	0,047	0,001	0,097	0,000	0,013	0,534	-0,019	0,719
Refinancing	0,030	0,020	-0,706	0,000	-0,280	0,000	-0,533	0,000
Property Type		0,000		0,000		0,000		0,751
Loan Purpose		0,000		0,009		0,101		0,557
Region		0,000		0,000		0,000		0,000
Loan Age		0,000		0,000		0,000		0,000

Table 21: Coefficient estimates and p-values Markov5(5) model.

5.3.3 Cross sectional analysis

The contingency tables for the cross sectional performance analysis of the three and five state Markov models are given in Table 22-23. The cross sectional analysis of the Markov models is slightly different from that of the MNL models due to the way in which the Markov models, or conditional MNL models, are estimated. For example, the Markov model conditional on partial prepayment is estimated by maintaining in the data set only those observations that are classified as partially prepaid. Estimating the cross sectional model performance is done by averaging the covariate information over the mortgages and predicting the state of a mortgage at the final date on which the mortgage is observed. Since all observations classified as partial prepayment are contained in the data set, no observations on partial prepayment remain to be predicted. This is accounted for in the tables by eliminating the states partial prepayment and delinquent for the Markov5(1) and Markov5(3) model, respectively.

	M3					**M5(1)**			
$\hat{Y}^3_{it}$ Y^3_{it}	**1**	**2**	**3**		$\hat{Y}^5_{it}$ Y^5_{it}	**1**	**3**	**4**	
1	2731	443	26	3200	**1**	3040	99	31	3170
2	84	5555	312	5951	**3**	1914	140	36	2090
3	20	611	211	842	**4**	0	0	0	0
	2835	6609	549	9993		4954	239	67	5260

Table 22: In-sample contingency table for Markov3 and Markov5(1) model.

The Markov3 models seems to perform reasonably well cross sectionally. The Markov5 model conditioned on partial prepayments slightly overestimates the number of full prepayments. This model fails to predict delinquent mortgages. Given the limited number of transitions from partial prepayment to delinquent, 102 (see Table 6), it is not surprising that this state is not accurately predicted by the model. Since the number of transitions from partial prepayment to default is zero this state is omitted from the table.

The contingency tables for the Markov5 models conditioned on delinquency status and contractual payment are provided in Table 23.

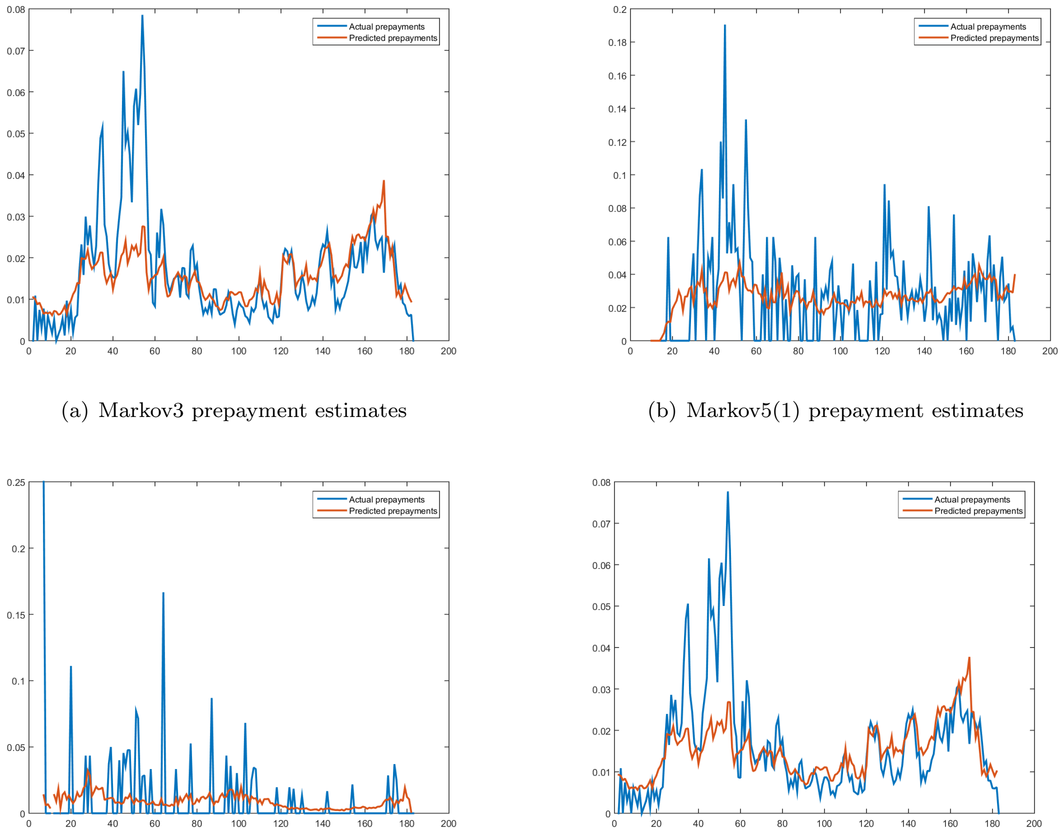
	M5(3)						**M5(5)**					
$\hat{Y}_{it}^5$ Y_{it}^5	**1**	**2**	**3**	**5**		$\hat{Y}_{it}^5$ Y_{it}^5	**1**	**2**	**3**	**4**	**5**	
1	870	56	26	5	957	**1**	237	222	135	3	8	605
2	231	48	5	0	284	**2**	369	1864	897	36	50	3216
3	97	4	8	1	110	**3**	19	209	4415	94	219	4956
5	59	2	1	6	68	**4**	2	13	43	5	7	70
						5	7	92	738	56	253	1146
	1257	110	40	12	1419		634	2400	6228	194	537	9993

Table 23: In-sample contingency table for Markov5(3) and Markov5(5) model.

The Markov5 model conditioned on delinquent payments slightly overestimates the default rate. The Markov5 model conditioned on contractual payments performs reasonably well and is comparable to the contingency table of the MNL5 model in Table 12.

5.3.4 Time series analysis

Figures 17(a)-17(d) show the time series performance of the Markov models over the sample period January 1999M until March 2014.



(a) Markov3 prepayment estimates

(b) Markov5(1) prepayment estimates

(c) Markov5(3) prepayment estimates

(d) Markov5(5) prepayment estimates

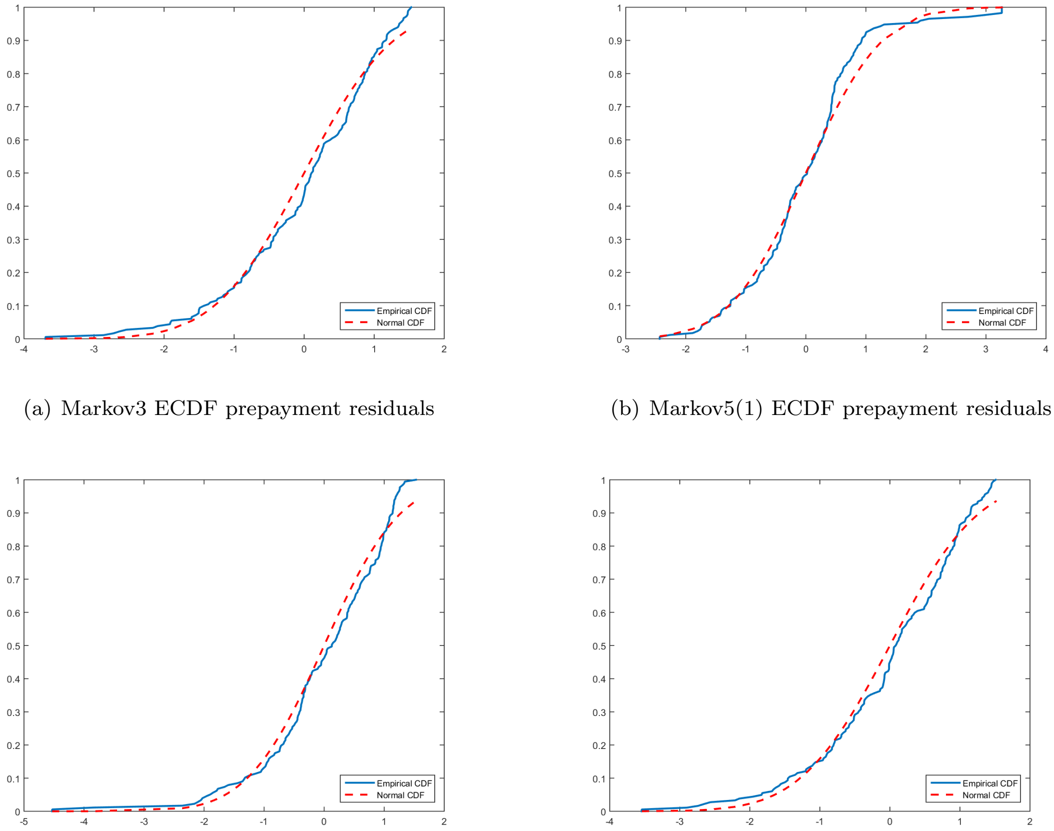
Figure 17: Time series analysis of Markov prepayment estimates.

In Figure 17(a) it can be seen that the three state Markov model is able to capture the general trend in prepayments over the sample period. The period November 2000 until October 2003 forms an exception. In this period prepayments are at their highest level which is underestimated by the Markov3 model. Figure 17(d) shows the predictions of the Markov5 model conditional on contractual payments. The two figures are very similar which is in line with the findings in the MNL3 and MNL5 model that the addition of intermediate states only influences the (full) prepayment predictions minorly.

Figures 17(b) and 17(c) present the prepayment estimates for the Markov5 models conditional on partial prepayments and delinquent payments. A first observation is that the actual prepayment rates fluctuate more extensively in these two figures compared to the other figures. This is due to the fact that these models are based on a less observations. The partial prepayment model is based on 10,596 observations (the number of partial prepayments in the sample) and the delinquent payment model on 5,421 observations (the number of delinquent payments in the sample). When stratifying prepayment rates per calender month, the prepayment rate is more volatile over the months when the number of observations is lower. Bearing this in mind, the Markov5(1) and Markov5(3) model appear to be able to capture the average trend in prepayments over the sample period.

5.3.5 Residual diagnostics

Figures 18(a) - 18(d) plot the empirical distribution of the prepayment residuals of the Markov models together with a normal distribution with the same mean and variance. The mean and variance of the residuals is given in Table 24.



(a) Markov3 ECDF prepayment residuals

(b) Markov5(1) ECDF prepayment residuals

(c) Markov5(3) ECDF prepayment residuals

(d) Markov5(5) ECDF prepayment residuals

Figure 18: Empirical and normal CDF of prepayment residuals in Markov models.

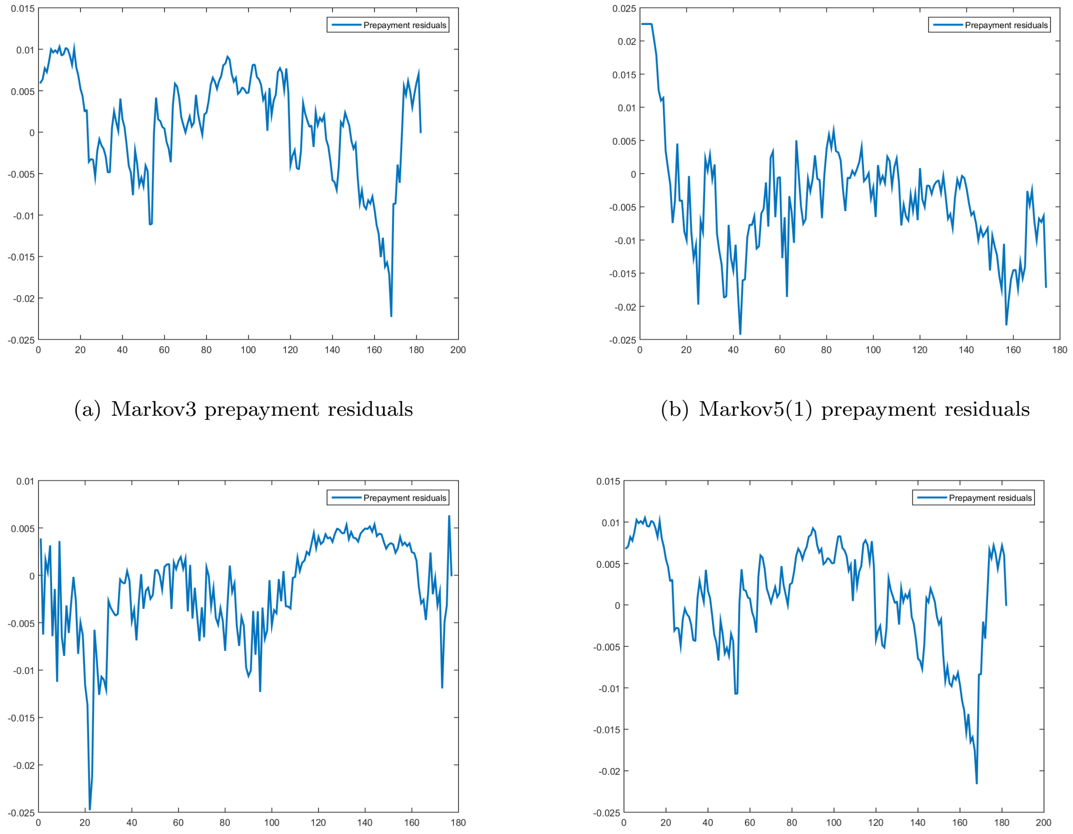
The ECDF of the prepayment residuals for all models, especially the Markov3 model and the Markov models conditioned on delinquency status and contractual payments closely resemble a normal distribution. This is confirmed by the Kolomogorov-Smirnov (KS) test statistics in Table 24 and indicates that the use of the CLT is justified. Since the Markov model conditional on partial prepayments is based on a lower number of observations it is no surprise that this residual ECDF is slightly further off from the normal CDF however, a normal distribution for these residuals can still not be rejected.

	$\bar{\mu}(\epsilon_{jt}) * 10^3$	$\hat{\sigma}^2(\epsilon_{jt}) * 10^3$	KS	p_{ttest}
Markov3	0,469	6,255	0.156	0.105
Markov5(1)	-4,517	8,442	0.238	0.000
Markov5(3)	-1,427	5,147	0.161	0.000
Markov5(5)	0,836	6,370	0.103	0.057

Table 24: Mean, variance, KS statistic and p-value zero mean hypothesis for prepayment residuals Markov models.

The presence of a zero mean in the residuals is rejected for the Markov3 model but can be confirmed for the Markov5(1) and Markov5(3) models at the one percent level and for the Markov5(5) model at the ten percent level, indicated by the p-value for the t-test in Table 24.

Figures 19(a) - 19(d) plot the prepayment residuals, ϵ_{tj} of the Markov models over the sample period.



(a) Markov3 prepayment residuals

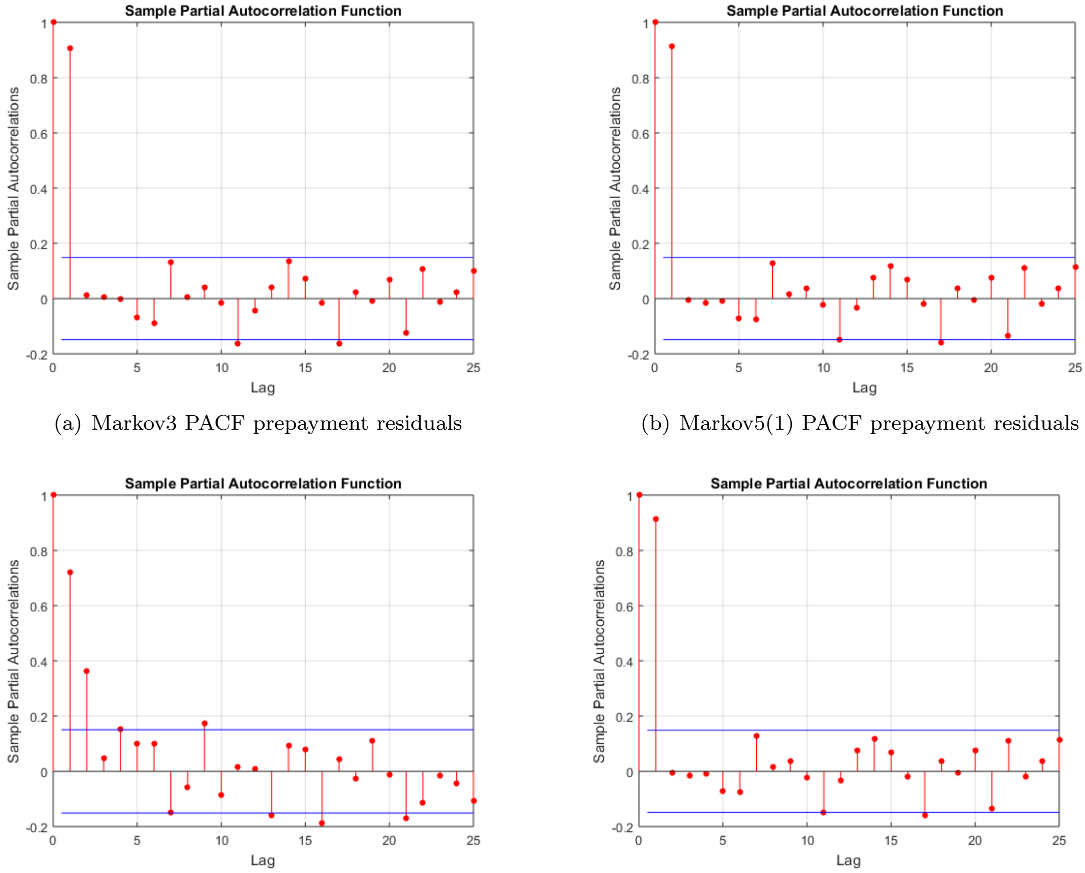
(b) Markov5(1) prepayment residuals

(c) Markov5(3) prepayment residuals

(d) Markov5(5) prepayment residuals

Figure 19: Time series analysis of Markov prepayment residuals.

In Figures 20(a) - 20(d) it can be seen that there is some time variation in the residuals. To determine the order of residual autocorrelation, the Partial Autocorrelation Functions (PACF) are plotted in Figures 14(a) and 14(b).



(a) Markov3 PACF prepayment residuals

(b) Markov5(1) PACF prepayment residuals

(c) Markov5(3) PACF prepayment residuals

(d) Markov5(5) PACF prepayment residuals

Figure 20: Partial Autocorrelation Functions of Markov prepayment residuals.

The prepayment residuals of both the Markov model exhibit significant first order autocorrelation. This is confirmed by the p-values of the Ljung-Box test on the AR(1) coefficient of the residuals under a $\sim \ \chi^2(1)$ in Table **??**.

	$\hat{\rho}_{\epsilon_1}$	$LB(1)$
Markov3	0.913	0.000
Markov5(1)	0.871	0.000
Markov5(3)	0.746	0.000
Markov5(5)	0.920	0.000

Table 25: Residual autocorrelation and Ljung-Box test statistic for Markov prepayment residuals.

The residuals of the Markov5 model conditioned on delinquency status have a lower first order autocorrelation compared to the other models, however the second order autocorrelation coefficient is also significant in this model.

Allthough the Markov models are constructed by conditioning on the current state in mortgage termination, there remains autocorrelation in the models which is not lower than the autocorrelation found in the MNL models. Similar remedies that were applied to the MNL models have been used for the Markov models, however these measures were only minorly effective and lead to less accurate forecasts and therefore they are not applied in the final models.

5.3.6 Forecasting

The contingency tables for the cross sectional out-of-sample performance for 200 mortgages for the Markov models is given in Tables 26 - 27.

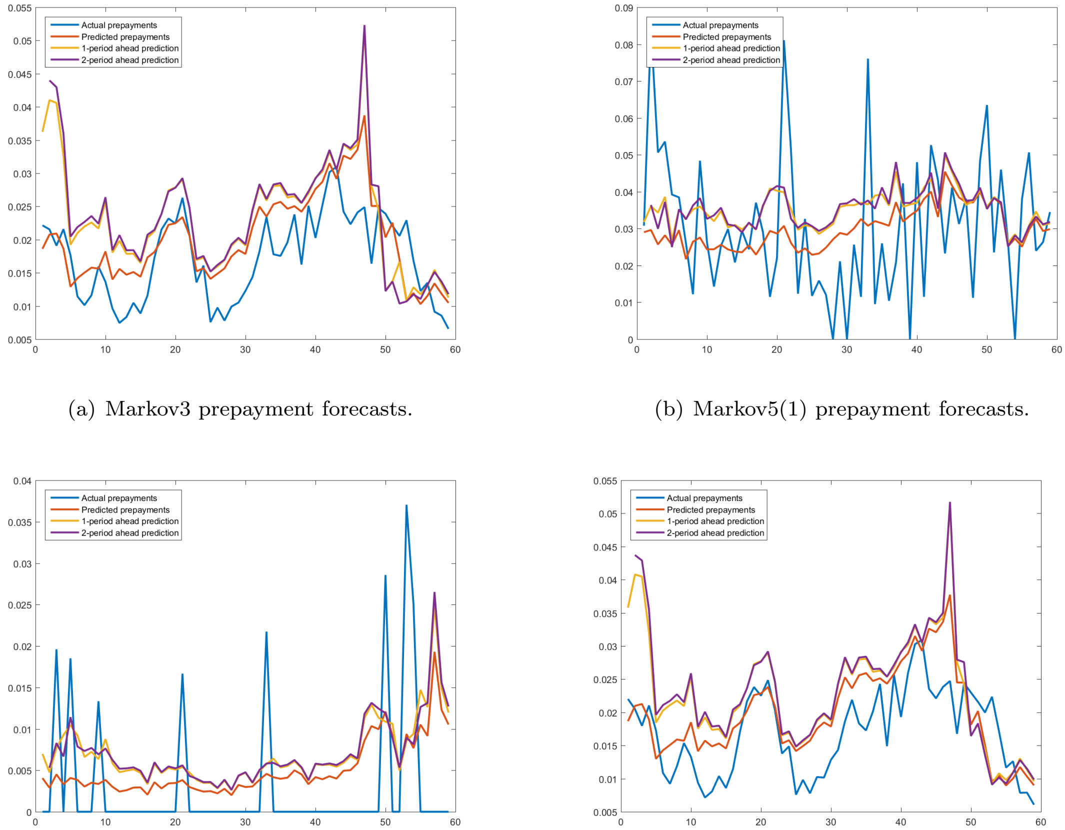
	M3					**M5(1)**			
$\hat{Y}^3_{it}$ Y^3_{it}	**1**	**2**	**3**		$\hat{Y}^5_{it}$ Y^5_{it}	**1**	**3**	**4**	
1	51	15	1	67	**1**	114	6	3	123
2	3	109	9	121	**3**	73	2	2	77
3	0	8	4	12	**4**	0	0	0	0
	54	132	14	200		187	8	5	200

Table 26: Out-of-sample contingency table for Markov3 and Markov5(1) model.

	M5(3)						**M5(5)**					
$\hat{Y}^5_{it}$ Y^5_{it}	**1**	**2**	**3**	**5**		$\hat{Y}^5_{it}$ Y^5_{it}	**1**	**2**	**3**	**4**	**5**	
1	132	11	8	5	171	**1**	8	4	4	0	0	16
2	14	4	0	0	22	**2**	10	30	22	0	0	62
3	5	1	0	0	7	**3**	0	3	91	3	7	104
5	0	0	0	0	0	**4**	0	0	0	0	0	0
						5	0	0	10	1	7	18
	151	16	8	5	200		18	37	127	4	14	200

Table 27: Out-of-sample contingency table for Markov5(3) and Markov5(5) model.

The time series out-of-sample performance of the models is again assessed by means of a one and two period out of time analysis using a sample split at January 2009. Figures 21(a) and 21(d) show the forecasted prepayment probabilities along with the actual prepayments and in-sample prepayment predictions for the one-step ahead forecasts of the Markov models.



(a) Markov3 prepayment forecasts.

(b) Markov5(1) prepayment forecasts.

(c) Markov5(3) prepayment forecasts.

(d) Markov5(5) prepayment forecasts.

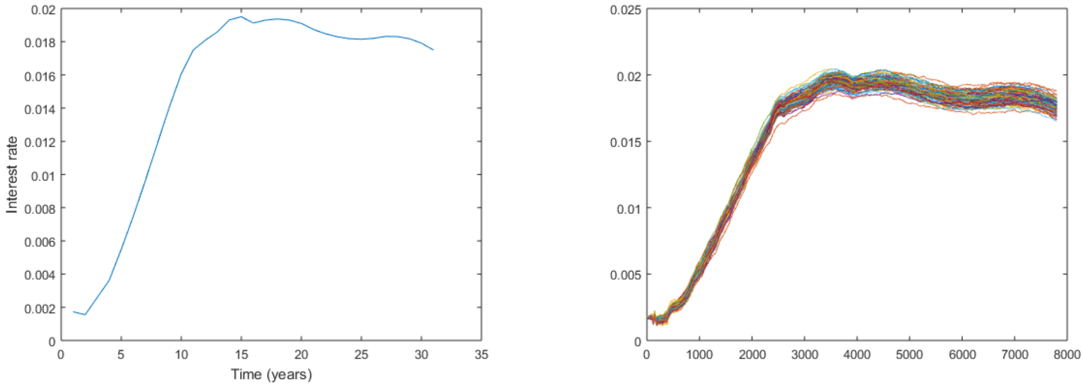
Figure 21: Markov models one -and two period ahead prepayment predictions January 2009 to March 2014.

From the figures it can be seen that the one step ahead prediction generally lies above the in-sample prediction from January 2009 to March 2013, after which the forecast drops below the in-sample prediction. The Markov3 forecast slightly overestimates the prepayment rate whereas the Markov5 models seem to be able to capture the general pattern in the prepayment rate.

5.4 Option theoretic model

The risk neutral refinancing incentive can be modelled as a European lookback put on the market mortgage rate. The strike price is fixed and equal to the contractual mortgage rate. The asset price is equal to the minimum value of the market mortgage rate over the lifetime of the mortgage, see Equation (3.23). The two stochastic processes that need to be modelled are the short rate and the market mortgage rate.

To model the short rate, the Hull-White model is calibrated on ATM swaption prices traded on the over-the-counter market (OtC), quoted in terms of Black volatilities and taken from Bloomberg. The interest rate term structure to which the model is fitted is the EURIBOR6M curve on 01-01-2015 provided by Bloomberg. This rate is chosen for calibration since US rates provided by Bloomberg are illiquid for longer maturities. Bloomberg does provide these rates but uses an extrapolation method that leads to inaccurate results. Since the maturities of the mortgages in this thesis is thirty years, it is important to obtain accurate rates for long maturities. The market implied term structure is given in Figure 22(a).



(a) Market implied term structure based on EURIBOR6M (01-01-2015).

(b) Simulated future paths for the instantaneous short rate for the Hull White model.

Figure 22: Market implied term structure and HW1f simulated short rate.

Since swaptions are quoted in terms of implied volatilities, Black's formula for swaptions has to be used to derive market prices, see Equation (3.41). Calibrating the HW1f model in a market consistent way implies finding values for a and σ_r for which the objective function in Equation (3.40) is minimized. The parameter estimates are given in Table 28.

a	$2.90 * 10^{-2}$
σ_r	$8.26 * 10^{-3}$

Table 28: Parameters for the Hull White model calibrated to swaption prices.

Next, future paths for the instantaneous short rate are simulated. The Euler scheme is used to discretize the stochastic differential equation. Figure 22(b) shows simulations for the short rate under the Hull White model (100 simulations). Although the interest rate level was low on the date of calibration, 01-01-2015, interest rates do not become negative. What is more, the simulated paths for the short rate are very similar. This is due to the fact that the variance of the short rate is very small, see Table 28.

Next, the market mortgage rate has to be modelled. Looking at Figure 4 it can be seen that the pattern of the 30FRM rate and the EURIBOR6M rate is similar over the sample period, 1999M01 until 2014M03.

ρ_{ma_t, r_t^f}	0.814
s_{ma-r^f}	3.200

Table 29: Correlation and spread between 30Y FRM and EURIBOR6M.

Indeed, it can be seen in Table 29 that the historical correlation between the market mortgage rate and the risk free rate is high. Under the assumption that the spread between the two rates is constant, the market mortgage rate can be modelled by the same stochastic process as the risk free rate. Table 29 also shows the average spread between the two rates over the selected sample period. This spread will be added to the simulated zero bond price process of the risk free rate to derive a stochastic process for the market mortgage rate.

The put option is ITM if the market mortgage rate for the remainder of the maturity of the mortgage is lower than the contractual mortgage rate in at least one month between the origination of the mortgage contract and its maturity (30 years). Since the strike price for the lookback option, e.g. the contractual mortgage rate, differs per mortgage the value of the put option is derived by taking the risk neutral expectation over 50 simulations for each mortgage. The large number of mortgages puts bounds on the number of simulations can can be conducted within reasonable computation time. One relation that should be observable is the fact that mortgages with a lower contractual mortgage rate should have a lower value for the put option.

The price of the put option will be used as an explanatory variable in the exogenous models as an alternative manner for capturing the refinancing incentive.

5.5 Models including put option

As an alternative indicator for the refinancing incentive in Equation (4.2), the MNL and Markov models have been estimated with the risk neutral price of the put option. This variable has been included by replacing the refinancing incentive in all specific models with the risk neutral put price (the refinancing incentive was significant in all models). The model estimates are given in Appendix F. In the MNL3 model the put price is insignificant for both prepayments and for defaults, and does not alter the significance or coefficients of other variables in the model. In the MNL5 model the put price is only significant for delinquent payments. In the Markov3 model the put price is significant for prepayments. The coefficients on the other variables are again similar to the model estimated with the refinancing incentive. In the Markov5(1) model the put price is significant for all categories and in the Markov5(3) model this variable is significant for prepayments at the five percent level. In the Markov5(5) model it is significant for all categories. Overall, the put price does not perform better in the models compared to the refinancing incentive. In the MNL models it performs worse and in the Markov models the performance is the same. This can be due to the fact that allthough the risk neutral put price is the theoretically optimal calculation of the incentive to refinance, the refinancing incentive defined in Equation 4.2 makes use of actual available information on the risk free rate per month per state and the discounted value of UPB payments. The put price can however be used as a proxy for the refinancing incentive if such actual information is unavailable. In the remainder of this thesis the models are be estimated with the refinancing incentive.

6 Model performance

6.1 In-sample performance

Using panel data, in-sample model performance is assessed based on (i) R^2; (ii) AIC and (iii) deviance. The cross sectional performance of the model is evaluated based on (iv) success rate for predicting prepayments; (v) the percentage of mortgages wrongly predicted as prepayments and (vi) the percentage of mortgages for which a prepayment is missed. The success rate is defined in Equation (3.53). The false alarm rate for prepayments in Equation (3.54) and the missed prepayment rate in Equation (3.55) are defined in relation to contractual payments. The time series performance of the models is based on (vii) unbiasedness via a hypothesis test of a zero mean in the forecast error; (viii) accuracy via the MSPE and (ix)-(x) efficiency by means of the coefficients of the Mincer Zarnovic regression.

The criteria of the general models (before variable selection) and specific models (after variable selection) are given in Table 30 for the MNL and Markov models.

	MNL3	MNL5	Markov3	Markov5(1)	Markov5(3)	Markov5(5)
(i)	R^2					
General	0,251	0,106	0,012	0,088	0,011	0,006
Specific	0,244	0,105	0,012	0,087	0,010	0,005
(ii)	AIC					
General	58,207	54,560	51,022	57,083	54,835	50,537
Specific	42,055	36,526	39,022	29,059	26,811	36,519
(iii)	Dev*10^{-4}					
General	7,220	21,161	7,215	0,751	0,938	18,545
Specific	7,216	21,178	7,216	0,768	0,952	18,638
(iv)	S_j					
General	0,848	0,826	0,853	0,636	0,275	0,802
Specific	0,856	0,794	0,841	0,586	0,200	0,709
(v)	F_j					
General	0,030	0,041	0,030	0,343	0,098	0,036
Specific	0,028	0,044	0,030	0,386	0,077	0,030
(vi)	M_j					
General	0,066	0,012	0,060	0,364	0,575	0,018
Specific	0,064	0,015	0,067	0,414	0,650	0,022
(vii)	E(ϵ_{tj})					
General	0,001	0,001	0,000	-0,005	-0,001	0,001
Specific	0,001	0,001	0,001	-0,004	-0,002	0,001
(viii)	MSPE*10^4					
General	0,027	0,030	0,472	7,309	7,914	1,096
Specific	0,032	0,032	1,146	7,589	7,935	1,112
(ix)	$\alpha_{0,MZ}$					
General	0,016	0,016	0,016	0,023	0,007	0,016
Specific	0,016	0,016	0,016	0,023	0,007	0,016
(x)	$\alpha_{1,MZ}$					
General	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
Specific	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000

Table 30: In-sample performance criteria for estimated models.

Based on Table 30 it can be seen that the models after variable selection have a better score on all criteria. In terms of R^2, the MNL3 model outperforms the other models. In terms of the AIC criterion, the Markov 5 models perform better than the MNL models which is mainly due to the fact that those models contain less variables. In terms of deviance the Markov models score better than the MNL models. The cross sectional performance indicators are considerably higher for the MNL models. The MNL3 model has a success rate of 84.8 percent while sending out only 3.0 percent false alarms and missing a prepayment for 6.6 percent of the mortgages. The performance of the MNL5 model is comparable. The Markov3 model and the Markov5 model conditioned on contractual payment also perform reasonably well. In therms of the time series performance criteria it can be seen that the expected value of the error term is close to zero for all models which means that the estimates are unbiased. In terms of efficiency both MNL models have the best score whereas the Markov5 model conditioned on partial prepayments appears the be the least efficient as indicated by the highest MSPE. Based on accuracy, all models are of comparable quality as the constant term in the Mincer Zarnovic regression is close to zero and the slope coefficient is equal to one.

Overall, based on the performance criteria discussed above as well as the contingency tables and the figures with time series performance, the MNL3 model has the highest in-sample performance. Since all models have a higher in-sample performance after variable selection is conducted, the out-of-sample performance is assessed only for the specific models.

6.2 Out-of-sample performance

6.2.1 Cross sectional out-of-sample

The contingency tables for the cross sectional out-of-sample performance have been provided in Tables 15 and 16 for the MNL models and in Tables 26 and 27. Table 31 contains the out-of-sample prepayment success rate, false alarm rate and missed rate for the sample of 200 mortgages.

		MNL3	MNL5	Markov3	Markov5(1)	Markov5(3)	Markov5(5)
(i)	S_j	0,818	0,785	0,826	0,250	0,250	0,717
(ii)	F_j	0,056	0,000	0,056	0,390	0,093	0,000
(iii)	M_j	0,106	0,023	0,114	0,750	0,688	0,031

Table 31: Cross sectional out-of-sample performance criteria for estimated models.

Table 31 shows that the three state models have the highest out-of-sample success rate. Moreover their false alarm rate and missed rate are low compared to the other models. The MNL3 model has a slightly lower success rate than the Markov3 model but this is compensated by a lower missed rate for the former. The MNL5 model has a relatively high success rate as well and a very low false alarm and missed prepayment rate. The Markov5(5) model scores relatively well. The performance of the other five state Markov models lacks behind. However, the Markov5(1) and Markov5(3) are based on a smaller set of observations. From the sample of 200 mortgages there are only 8 and 16 prepayments that can be predicted by these models, respectively. Therefore the weight of individual observations is large in these rates.

6.2.2 Time series out-of-sample

The one step ahead prepayment predictions have been given in Figures 15(a)-15(b) and Figures 21(a)-21(d). As explained in Section 3.5.3, forecasting performance is assessed in terms of (i) unbiasedness; (ii) accuracy and (iii) efficiency. Table 32 gives (i) the expected value of the forecast error; (ii) the MSPE and (iii) the Mincer-Zarnovic (MZ) coefficients for the one- and two- step ahead forecasts for all models (after variable selection).

	MNL3	MNL5	Markov3	Markov5(1)	Markov5(3)	Markov5(5)
				$\mathrm{E}(\epsilon_{tj})$		
One-period	-0.001	-0.001	-0.007	-0.006	0.002	-0.007
Two-period	-0.001	-0.001	-0.007	-0.006	0.002	-0.007
				MSPE*10^4		
One-period	0.048	0.016	0.880	3.539	0.912	0.850
Two-period	0.130	1.075	1.041	3.696	0.914	0.983
				$\alpha_{0,MZ}$		
One-period	0.016	0.016	0.016	0.032	0.010	0.016
Two-period	0.016	0.002	0.017	0.032	0.010	0.016
				$\alpha_{1,MZ}$		
One-period	-1.012	-1.005	-1.009	-1.081	-1.088	-1.010
Two-period	-1.024	-0.292	-1.014	-1.067	-1.082	-1.015

Table 32: IOut-of-sample performance criteria for estimated models.

Table 32 shows that for the majority of the models the expected value of the one step ahead residuals is slightly negative. This means that the models overestimate the prepayment rate, which was visible in the time series figures. An asymptotoc t-test indicates that the model forecasts are unbiased. Based on accuracy the models rank the same in-sample and out-of-sample. The MNL models again have the lowest MSPE, with the MNL5 model having the lowest MSPE. The Markov5 models are the least efficient. In terms of accuracy, the models show more divergence out-of-sample compared to in-sample. The MNL5 model has a slightly better score than the MNL3 model for this criterion.

All in all, the MNL models have the highest score on the selected criteria, both in-sample and out of sample, both cross-sectionally as well as time series wise. Furthermore, the performance for predicting (full) prepayment rates of the two MNL models is comparable. This indicates that the inclusion of partial prepayments as an additional state in the MNL model does not lead to better results. Therefore, the model that is selected as best performer is the MNL3 model. A final test that is applied to this model is a test on parameter stability.

6.3 Tests for parameter stability

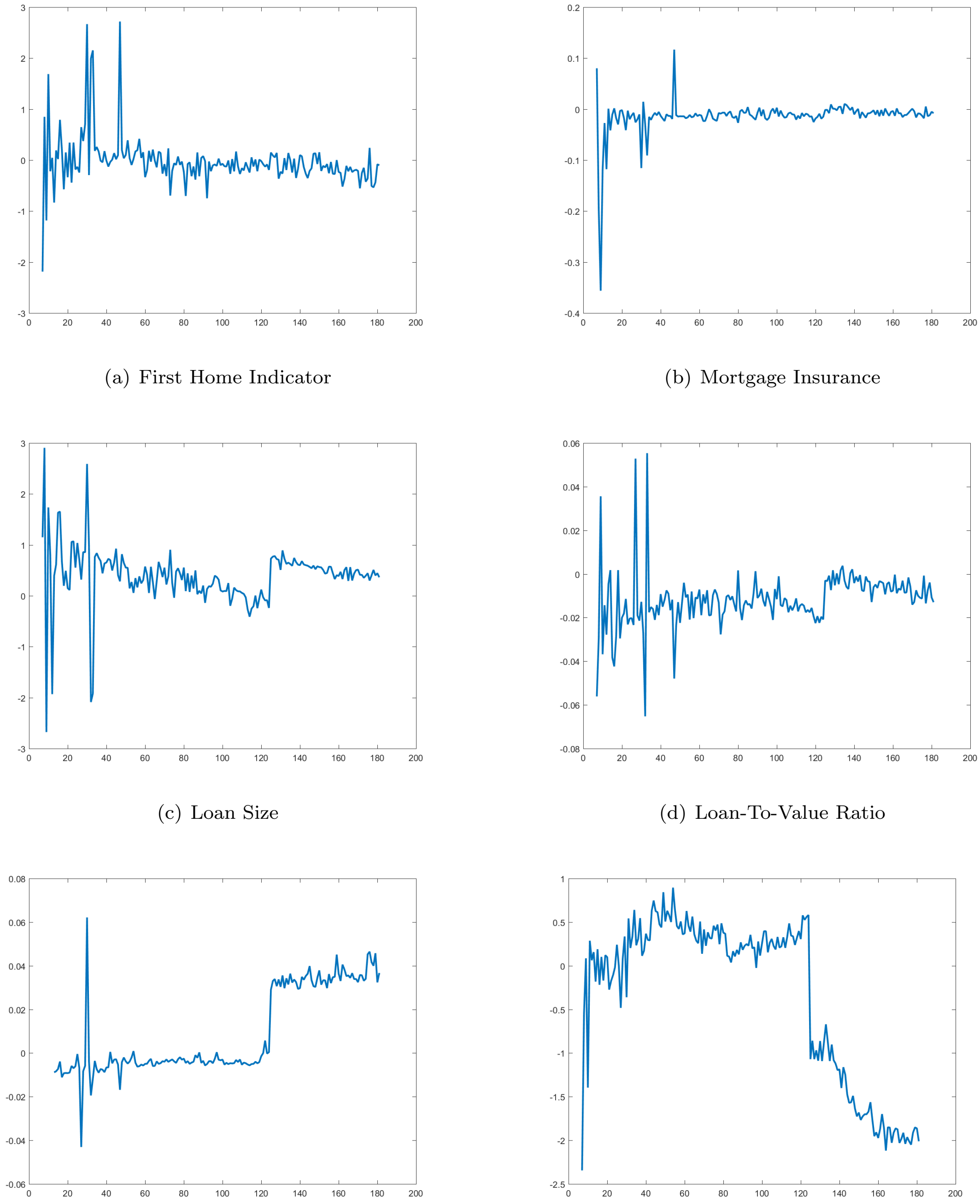
The data set spans the period January 1999 until March 2014 and therefore contains data on the 2008 credit crisis. The sixteen years of data and can be split into nine pre-crisis years and six post crisis years. Since prepayment models before, during and after a crisis are likely to be different, it is interesting to investigate whether the estimated parameters are stable over time or whether the crisis constitutes a structural break in the prepayment models. To test for parameter stability the CUSUMSQ test will be conducted. The advantage of the CUSUMSQ test over Quandt's Likelihood Ratio test is that the former is a lot shorter in computation time. The reason for this is that the QLR test requires the estimation of twice the number of coefficients which is time consuming given the sizable data set at hand. See Appendix G for the specification of the QLR test. The CUSUM test requires the estimation of time varying coefficients. To this end the MNL model is estimated T times using observations from months $t = 1, ..., 182$. This procedure was initially applied to the sample of 10,000 mortgages. However, for some months teh number of observations was limited int his case. Therefore, the final CUSUMSQ test is applied to the full data set. The test on parameter stability is applied to the model which is classified as best performer according to the previous sections, namely the MNL3 model.

6.3.1 Estimation of time varying coefficients

The CUSUMSQ test is conducted using the full data set and the variables contained in the specific MNL3 model. These variables can be found in Table 9. The panel data set is restructured by calender month and the coefficients are estimated monthly for each month between January 1999 to March 2013 (183 estimates). In each month it is tested whether the explanatory variables contain values and exhibit a correlation of below |1| with any of the other explanatory variables. If either of the two is not the case, the variables are removed from the model. Therefore, the number of variables in the model can differ per calender month. This mostly holds for a subset of dummy variables for categorical risk drivers such as property type and loan purpose in the beginning or at the end of the sample period. Furthermore, loan age dummies often drop out of the model in the beginning of the sample period due to the structure of the data set.

This requirement on the included variables also constituted a main reason to perform the CUSUMSQ tests on the full data set. The CUSUMSQ tests have been applied to the sample of 10,000 mortgages but lead to issues of missing covariate information per month or too limited variation in covariates in a certain month. Moreover if in one month no prepayment or default occurred, the MNL model could not be estimated at all.

Figures 23(a) - 23(f) show the coefficient estimates for prepayments per calender month for the full sample for the non-categorical variables of the MNL3 model.



(a) First Home Indicator

(b) Mortgage Insurance

(c) Loan Size

(d) Loan-To-Value Ratio

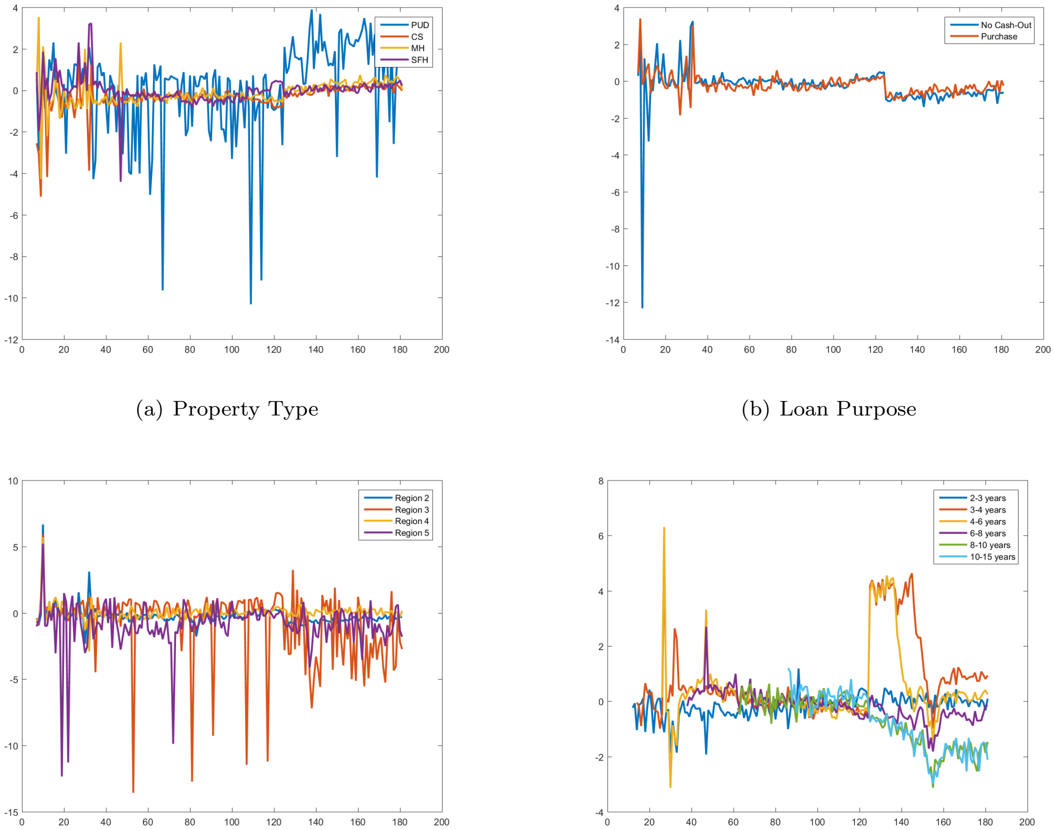
(e) House Price Index

(f) Refinancing Incentive

Figure 23: Time varying coefficient estimates non-categorical variables in MNL3 model (full data set).

The figures show that in general the estimated parameters are most fluctuative in the beginning of the sample period. One reason for this initial instability is fewer observations in the beginning of the sample. Parameter estimates for the First Home Indicator, Mortgage Insurance and Refinancing Incentive seem to be relatively stable over the remainder of the sample. For the variables Loan Size, LTV and House Price Index, a sharp rise around October 2009 is visible. After this point these variables have a larger effect on prepayment rates. This effect is most profound for the House Price Index. It appears that after the crisis of 2008 an increase in the House Price Index has a large positive effect on the prepayment rate compared to the pre-crisis period. The refinancing incentive shows a sharp drop just before October 2009. The coefficient becomes more negative. This indicates that a favorable market mortgage rate, e.g. lower than the contractual mortgage rate, after this period leads to a more pronounced refinancing incentive compared to the pre-crisis period.

Figures 24(a) - 24(d) show the time varying prepayment coefficient estimates for the categorical variables of the MNL3 model.



(a) Property Type

(b) Loan Purpose

(c) Region

(d) Loan Age

Figure 24: Time varying coefficient estimates categorical variables in MNL3 model (full data set).

The coefficients for Property Type fluctuate the most in the beginning of the sample. The estimates for the property types Cooperative Share, Manufactured Housing and Single Family Home stabilize over time while the coefficient for Planned Unit Developments remains very volatile. The coefficients for Loan Purpose are relatively stable while the estimates for Region fluactuate quite a bit. Especially for the second and fifth region. From the Loan Age variables the spikes of loan ages between three and six years stand out. Apparently, loans of this loan age have an increasing probability of being prepaid in the period near the end of 2009.

All in all, the coefficient estimates for prepayments are quite fluctuatuve during the sample period. Hence it is expected that the parameters are not stable over time. This will be tested more formally with the CUSUMSQ test.

6.3.2 CUSUMSQ Test

The CUSUMSQ test is based on the recursive residuals given in Equation (3.18). Applied to the logit model, these residuals are adapted slightly. Since the aim of the thesis is to determine a prepayment model, the test will be applied to prepayment residuals only. The recursive prepayment residuals are defined as

$$w^*_{t,j} = \frac{\epsilon^*_{tj}}{\sqrt{1 + x_t(X'_{t-1}X_{t-1})^{-1}x'_t}} \tag{6.1}$$

in which ϵ^*_{tj} denotes the MNL residual for state j and is defined in Equation (3.57).

The CUSUMSQ test statistic, S_t, is plotted in Figure 6.3.2 for $T = 182$ and $k = 28$, the average number of parameters in the time series models.

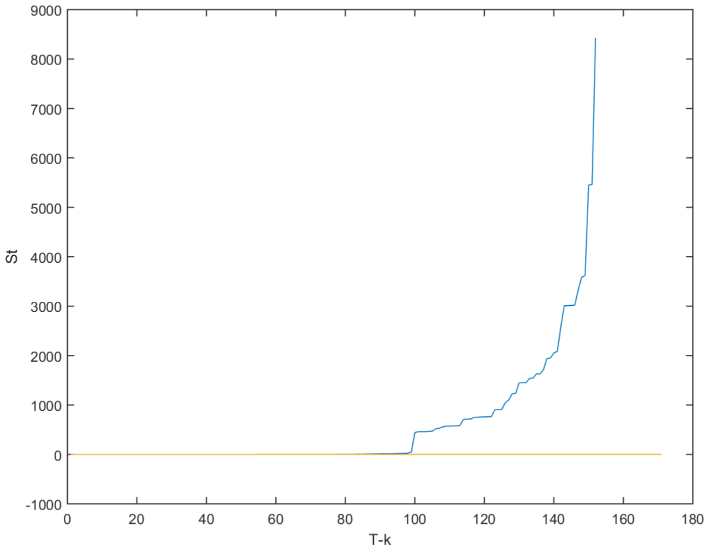


Figure 25: CUSUMSQ test MNL3 model.

The CUSUMSQ plot indicates the presence of a structural break at $T - k = 100$, which corresponds to October 2009. This timing is in line with the spikes in some of the parameter estimates in this period, visible mainly in Figures 23(c) through 23(c) and Figure 24(d). The presence of a break point in prepayment rates is also visible in Figure 5, in which prepayment rates are plotted over time. It appears that prior to October 2009 prepayment rates were fluctuating around a higher mean compared to post October 2009. This break date can also be linked to the credit crisis of 2008. Prior to this crisis economic growth enabled unscheduled return of the mortgage principal while after the crisis prepayment rates are vastly lower. Indeed Figure 5 shows that after the crisis prepayment rates fluctuate around a lower mean.

6.3.3 MNL3 model with structural break

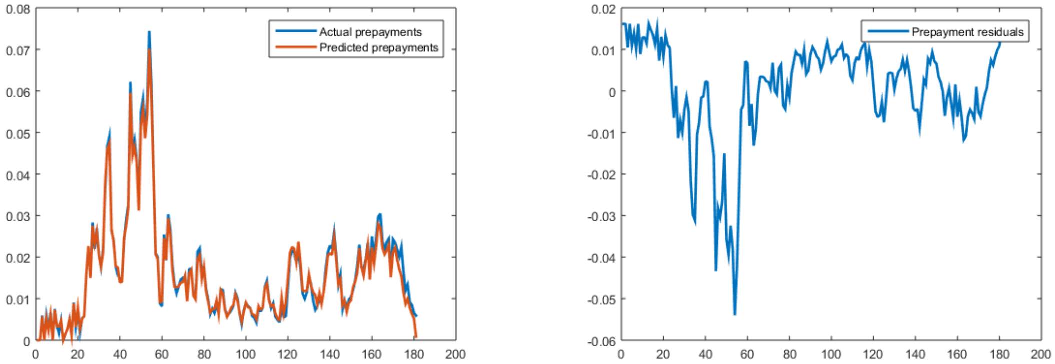
The MNL3 model will consequently be estimated by allowing for different parameters before and after October 2009. To this end the MNL model will be estimated with an indicator variable, according to Equation (3.21). The number of observations before October 2009 denotes 237,198 and is slightly higher than the number of observations after this time, which equals 172,121. The model estimates are given in Table 33. The individual estimates for the categorical variables are given in Table 49.

	Prior	**Crisis**			**Post**	**Crisis**		
	Prep		Default		Prep		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-13,407	0,000	-18,287	0,000				
Firsthome	0,053	0,928	0,240	0,682	-0,137	0,475	-0,139	0,612
Mortgage insurance	-0,031	0,106	-0,014	0,476	-0,016	0,017	-0,018	0,045
Loansize	0,074	0,853	0,399	0,315	0,349	0,001	0,119	0,437
LTV	0,020	0,131	0,049	0,001	0,003	0,568	0,048	0,000
Houseprice	0,004	0,435	0,011	0,022	0,013	0,002	0,013	0,016
Refinancing	0,039	0,862	-0,258	0,246	-1,597	0,000	-1,713	0,000
Loan Age		0,000		0,000		0,000		0,000
Property Type		0,000		0,000		0,000		0,000
Loan Purpose		0,000		0,000		0,000		0,000
Region		0,000		0,000		0,000		0,000

Table 33: Coefficient estimates and p-values MNL3 model with structural break.

The estimates show that while the majority of the variables are significant in the post crisis period, they are not significant pre-crisis. This indicates that pre and post crisis prepayment rates differ in terms of the risk drivers by which they can be modelled. Furthermore, since the variables that were selected in the original MNL3 model are the variables that were significant in the post crisis period it seems that the post crisis period largely influenced the variable selection procedure even though this period comprises less observations. The model can be extended by allowing different variables in the model depending on the regime in which the model is.

The performance of the model over time is given in Figure 26(a) and the prepayment residuals in 26(b).



(a) MNL3 model estimates.

(b) MNL3 model prepayment residuals.

Figure 26: MNL3 model estimates with structural break at October 2009.

With an R^2 of 0.256 the model scorers slightly better than the MNL3 model without structural breaks. The MSPE of the model is also low compared to the other models, namely $2.132 * 10^{-6}$. However this comes at the expense of a sharp increase in the AIC, which amounts 100.317. This is no surprise given that the number of parameter estimates has doubled. The residual AR(1) coefficient has decreased to 0.881 but remains relatively high.

7 Conclusion

In this thesis four types of prepayment models are estimated for a mortgage portfolio of Freddie Mac. Three of these models are widely used in the prepayment literature, namely the option theoretic model, the multinomial logit model (MNL) and the competing risk model. A Markov model has been introduced as a new method for modelling prepayment rates. The aim of this thesis is to determine which of these models is the best performer for predicting full prepayment rates both in-sample and out-of-sample, whereby out-of-sample performance is assessed cross sectionally as well as time series wise.

The option theoretic model is constructed as a lookback put option on the mortgage. This is an interesting from a theoretical standpoint however its major drawback lies in the inability of incorporating the behavioural risk that resides in prepayment decisions. Such behavioural risk arises from the fact that mortgagees are unaware of optimal prepayment options as well as from diverging borrower specific attributes such as the Debt-to-Income ratio, FICO score, region in which the property is located and the divorce rate. To incorporate the fact that the optimal prepayment decision merely constitutes one indicator for prepayment, the risk neutral put price has been included in exogenous models as an explanatory variable. Exogenous or empirical models aim to estimate prepayment rates according to a number of borrower specific, loan specific and macro economic variables. Including the put price as explanatory variable is a good alternative to incorporate the refinancing incentive however leads to less accurate predictions compared to using a straightforward variable for the refinancing incentive.

The most popular exogenous model in prepayment modelling is the MNL model. Two MNL models are estimated. The three state MNL model predicts mortgage termination based on the competing risks full prepayment and default. The five state MNL model incorporates the transient states partial prepayment and delinquent payments. The predictions of the MNL models with three and five states were fairly similar. The addition of partial prepayments is therefore not necessary for predicting full prepayment rates. An indication for this was also provided during the variable selection procedure, in which variables that were significant for prepayments were not significant for partial prepayments and vice versa. The main risk drivers for prepayments are the refinancing incentive, house prices, the Loan-to-Value ratio, the unemployment rate and loan size while these variables appear to be insignificant for partial prepayments in the majority of the models. Overall the MNL predictions performed very well. Out of sample forecasts scored well in terms of accuracy and efficiency. Also the cross sectional analysis of model performance revealed high success rates and low false alarm -and missed prepayment rates for the MNL models. The major drawback of MNL models applied to panel data is the implicit assumption of independent consecutive observations. This drawback became visible when assessing time series model performance. However, given the fact that prepayment rates itself as well as well as the main macro-economic variables affecting them show strong dependence over time autocorrelation is inherent.

The competing risk model is also evaluated and it is shown that for small values of the baseline hazard the similarity of this model and the MNL model is considerable. The major disadvantage of the competing risk model is its limited ability to include time varying explanatory variables. Given the availability of a panel data set and the objective of this thesis to perform out of time analyses, the added value of the competing risk model does not extend beyond the estimation of the baseline survival function. This survival function has shown a relatively steep prepayment survival curve and indicated that the effect of loan age on prepayment rates is not linear. To incorporate this, loan age is included in the MNL model as a categorical variable.

The transition probabilities of the Markov models are estimated using covariate information and by conditioning on the transient states of mortgage payment. The three state Markov model is obtained by conditioning on contractual payment. The five state Markov models are estimated by means of three conditional MNL models, conditioned on partial prepayment, delinquent payment and contractual payment. Allthough the Markov models account for dependence between observations, this appeared to be insufficient for incorporating the majority of the autocorrelation present in prepayment rates.

The performance of the three exogenous models has been evaluated using a wide set of performance measures. The panel data set criteria include the R^2, the Akaike information criterion and the deviance. Cross sectional performance is evaluated with contingency tables that compare predicted and actual state realizations across mortgages. Time series wise, the models are evaluated based on unbiasedness, accuracy and efficiency. The distribution of the residuals is assessed as well as its autocorrelation properties. Based on the performance assessment of the models, the three state multinomial logit model outperforms the other models cross sectional wise and time series wise. Even though this model proved to be incapable of incorporating all time series dependence present in prepayment rates, its in-sample and out-of-sample predictions are accurate. Allthough estimation of the Markov models requires a factor three more parameters to be estimated (one model per transient state), the performance of these models lacks behind the performance of the multinomial logit model. This is mainly caused by the limited number of observations for, especially, partial prepayment and delinquent payments. In the cross sectional analyses this lead to success rates, false alarm rates and missed prepayment rates that were highly influenced by the prediction of only a few prepayment observations. This was already an issue when evaluating in-sample model performance but was exemplified during the out-of-sample analysis in which some transitions could not be predicted at all. In the time series analysis the limited number of observations lead to a sharp fluctuation of the prepayment rate over calender months, a feature that could not be captured by the Markov models.

Concluding, the multinomial logit model is the most appropriate model for determining prepayment rates both time series wise and cross-sectional wise as well as in-sample and out-of-sample. The competing risk model was insufficient for incorporating time varying covariates. Since macro economic variables are an important indicator for prepayment rates, this model is not a well-suited prepayment model. Allthough the Markov model theoretically seems a good candidate for a prepayment model, it suffers from some practical drawbacks. The main weakness of this model lies in the fact that it produces less reliable estimates in case the number of observations for a certain transition is limited. Given the size of the mortgage market mortgage level data can in general be obtained in large quantities. However, the requirement of a lot of data for estimation comes at the expense of a sharp increase in computation time. Taking this into consideration, Markov models can therefore better be applied to systems in which transition probabilities are more evenly spread across the state space.

Since the data set includes the credit crisis of 2008, which is likely to have a large impact on prepayments rates, the parameter estimates of the three state MNL model are tested on stability. A structural break is found shortly after the crisis. Prior to the crisis prepayment rates were significantly higher than in the post crisis period. This is incorporated by allowing for different coefficients in these periods. Since pre and post crisis prepayment rates appear to differ in terms of the risk drivers by which they can be modelled, the prepayment model can be extended by allowing for different variables in the model depending on the regime in which the model is.

A limitation of this research is the inability of the MNL models to incorporate dependency between consecutive observations in a panel data set. Particularly in a prepayment model this constitutes a challenge. Future research could focus on methods of incorporating the strong dependence between observations.

A Similarity MNL and competing risk likelihood

Starting from the cause specific hazard rate defined in Equation (3.6),

$$\lambda_j(t,x) = \lambda_{j0}(t) \exp\left(-X'_{it}\beta_j\right) \tag{A.1}$$

this can be rewritten as

$$\lambda_j(t,x) = \exp\left(\log\left(\lambda_{j0}(t)\right) + X'_{it}\beta_j\right). \tag{A.2}$$

Defining $z_{itj} = \log\left(\lambda_{j0}(t)\right) + X'_{it}\beta_j)$, the conditional probability of mortgage termination due to cause j is given by

$$\frac{\exp z_{itj}}{\sum\limits_j \exp z_{itj}}. \tag{A.3}$$

Comparing this to the MNL probability given in Equation (3.2) it can be seen that the two models are comparable when the baseline hazard rate $\lambda_{j0}(t)$ is small.

B Summary statistics full data set

Variable	Mean	St. Dev.	Min	Max
Loan Age	32.361	28.601	1	181
FICO score	729.653	54.576	360	850
Firsthome	0.139	0.345	0	1
Mortgage insurance	5.032	10.444	0	52
DTI	34.164	11.796	1	75
Loansize	0.915	0.509	0.048	7.101
LTV	72.816	16.025	6	104
Penalty	0.002	0.041	0	1
Unemployment	6.156	2.099	2.100	14.900
Divorcerate	4.046	0.896	1.700	9.900
Houseprice	155.734	26.367	100.590	206.670
Refinancing	-0.605	0.990	-7.660	6.560
Property Type Condo	0.007	0.708	0	1
Property Type Planned Unit Development	0.271	0.444	0	1
Property Type Cooperative Share	0.303	0.459	0	1
Property Type Manufactured Housing	0.253	0.435	0	1
Property Type Single Family Home	0.181	0.385	0	1
Purpose Cash-Out	0.652	0.476	0	1
Purpose No Cash-out Refinance	0.100	0.300	0	1
Purpose Purchase	0.248	0.432	0	1
Region 1	0.074	0.263	0	1
Region 2	0.000	0.022	0	1
Region 3	0.145	0.352	0	1
Region 4	0.008	0.090	0	1
Region 5	0.772	0.420	0	1
Loan Age below 1Y	0.294	0.456	0	1
Loan Age 2-3Y	0.212	0.409	0	1
Loan Age 3-4Y	0.151	0.358	0	1
Loan Age 4-6Y	0.183	0.386	0	1
Loan Age 6-8Y	0.094	0.291	0	1
Loan Age 8-10Y	0.045	0.207	0	1
Loan Age 10-15Y	0.021	0.144	0	1

Table 34: Summary statistics of dependent variables for full data set (29,932,667 obs).

C Estimated general models

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
C	-13,121	0,000	-9,920	0,000
FICO score	-0,002	0,060	-0,016	0,000
Firsthome	-0,477	0,003	-0,346	0,110
Mortgage insurance	-0,019	0,001	-0,021	0,005
DTI	-0,002	0,721	0,038	0,000
Loansize	-0,163	0,070	-0,194	0,136
LTV	0,015	0,000	0,048	0,000
Penalty	-1,596	0,656	-0,019	0,996
Unemployment	0,087	0,054	0,320	0,000
Divorcerate	-0,230	0,076	-0,361	0,006
Houseprice	0,025	0,000	0,027	0,000
Refinancing	-1,376	0,000	-1,380	0,000
Dummy Loan ID	20,400	0,000	18,258	0,000
Dummy FICO	-1,591	0,348	-15,574	0,000
Dummy Firsthome	-1,041	0,000	-1,401	0,000
Dummy Mortgage Insurance	2,904	0,000	2,343	0,001
Dummy DTI	0,210	0,620	2,117	0,000
Dummy Penalty	1,802	0,004	1,987	0,005
Dummy Unempl	-4,066	0,000	-2,204	0,000
Dummy divorcerate	-1,112	0,061	-1,320	0,029
Dummy Houseprice	-2,237	0,002	-2,675	0,001
Dummy Refinancing	2,831	0,000	3,668	0,000
Propertytype Planned Unit Development	-0,092	0,510	0,920	0,000
Propertytype Cooperative Share	0,298	0,024	0,897	0,000
Propertytype Manufactured Housing	-0,380	0,002	-0,337	0,055
Propertytype Single Family Home	-4,478	0,000	-4,885	0,000
Purpose No Cash-out Refinance	-2,078	0,000	-2,137	0,000
Purpose Purchase	-2,115	0,000	-1,508	0,000
Region 2	-2,155	0,279	-5,537	0,234
Region 3	-0,202	0,374	-0,519	0,082
Region 4	-0,500	0,337	0,760	0,207
Region 5	0,010	0,959	-0,248	0,315
Loan Age 2-3Y	0,330	0,036	0,554	0,017
Loan Age 3-4Y	-0,517	0,003	0,278	0,246
Loan Age 4-6Y	-1,164	0,000	-0,263	0,263
Loan Age 6-8Y	-2,235	0,000	-1,401	0,000
Loan Age 8-10Y	-2,614	0,000	-2,115	0,000
Loan Age 10-15Y	-3,307	0,000	-3,096	0,000

Table 35: Coefficient estimates and p-values MNL3 model before variable selection.

	Pa.Prep		**Prep.**		**Delinq.**		**Default**	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-4,716	0,000	-14,953	0,000	3,355	0,000	-11,811	0,000
FICO score	-0,001	0,000	-0,001	0,234	-0,016	0,000	-0,015	0,000
Firsthome	0,023	0,539	-0,249	0,093	-0,160	0,003	-0,120	0,580
Mortgage ins	0,003	0,015	-0,015	0,004	0,000	0,924	-0,018	0,018
DTI	0,000	0,877	-0,005	0,255	0,020	0,000	0,035	0,000
Loansize	0,059	0,013	-0,061	0,476	-0,093	0,006	-0,103	0,438
LTV	-0,001	0,514	0,010	0,005	0,014	0,000	0,044	0,000
Penalty	-0,590	0,127	-1,702	0,549	-1,426	0,156	-0,104	0,972
Unemployment	0,022	0,000	0,166	0,000	0,059	0,000	0,405	0,000
Divorcerate	0,063	0,000	-0,227	0,020	-0,041	0,017	-0,356	0,001
Houseprice	-0,001	0,164	0,019	0,000	0,004	0,000	0,021	0,000
Refinancing	-0,110	0,000	-1,153	0,000	-0,371	0,000	-1,178	0,000
Dummy Loan ID	5,031	0,000	18,645	0,000	4,008	0,000	16,726	0,000
Dummy FICO	-0,449	0,084	-1,124	0,358	-10,761	0,000	-16,240	0,002
Dummy Firsthome	-0,017	0,563	-0,867	0,000	-0,117	0,002	-1,232	0,000
Dummy Insurance	0,246	0,010	2,408	0,000	0,531	0,000	1,864	0,018
Dummy DTI	-0,078	0,336	-0,124	0,752	0,999	0,000	1,830	0,000
Dummy Penalty	-0,008	0,963	1,650	0,026	-0,040	0,829	1,890	0,025
Dummy Unempl	0,021	0,740	-2,643	0,000	-0,124	0,228	-0,716	0,088
Dummy divorcerate	0,303	0,000	-0,877	0,048	-0,269	0,001	-1,082	0,030
Dummy Houseprice	0,180	0,016	-0,215	0,683	0,455	0,000	-0,599	0,336
Dummy Refinancing	-0,114	0,666	2,374	0,000	1,857	0,000	3,332	0,000
Propertytype PUD	-0,041	0,178	-0,064	0,628	0,276	0,000	0,963	0,000
Propertytype CS	-0,049	0,107	0,224	0,076	0,332	0,000	0,842	0,000
Propertytype MH	0,041	0,234	-0,231	0,066	-0,070	0,131	-0,189	0,307
Propertytype SFH	0,159	0,000	-2,906	0,000	-0,035	0,481	-3,329	0,000
Purpose No Cash	0,068	0,124	-1,603	0,000	-0,092	0,111	-1,691	0,000
Purpose Purchase	0,030	0,387	-1,634	0,000	0,257	0,000	-1,046	0,000
Region 2	-0,428	0,241	-1,335	0,370	-6,867	0,454	-5,800	0,426
Region 3	0,074	0,160	0,074	0,729	0,330	0,000	-0,257	0,391
Region 4	-0,074	0,566	-0,160	0,752	0,289	0,027	1,079	0,081
Region 5	0,125	0,005	0,229	0,212	0,283	0,000	-0,039	0,876
Loan Age 2-3Y	1,172	0,000	0,882	0,000	0,540	0,000	1,153	0,000
Loan Age 3-4Y	1,316	0,000	0,538	0,001	0,633	0,000	1,378	0,000
Loan Age 4-6Y	1,452	0,000	0,094	0,558	0,616	0,000	1,035	0,000
Loan Age 6-8Y	1,637	0,000	-0,217	0,272	0,421	0,000	0,615	0,029
Loan Age 8-10Y	1,735	0,000	-0,562	0,010	0,181	0,014	-0,058	0,859
Loan Age 10-15Y	2,049	0,000	-0,898	0,000	-0,081	0,387	-0,707	0,080

Table 36: Coefficient estimates and p-values MNL5 model before variable selection.

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
C	-5,418	0,000	-4,431	0,000
FICO score	0,001	0,000	-0,011	0,000
Firsthome	-0,110	0,012	-0,020	0,898
Mortgage ins	0,000	0,886	-0,003	0,600
DTI	-0,007	0,000	0,029	0,000
Loansize	0,464	0,000	0,393	0,000
LTV	-0,003	0,002	0,029	0,000
Penalty	-0,648	0,266	0,991	0,326
Unemployment	-0,096	0,000	0,077	0,001
Divorcerate	0,104	0,000	-0,019	0,714
Houseprice	-0,001	0,183	0,001	0,697
Refinancing	-0,685	0,000	-0,499	0,000
Dummy FICO	1,318	0,000	-24,371	0,992
Dummy Firsthome	-0,041	0,218	-0,330	0,012
Dummy Insurance	0,694	0,000	0,314	0,539
Dummy DTI	-0,387	0,000	1,427	0,000
Dummy Penalty	0,093	0,567	0,617	0,185
Dummy Unempl	-0,289	0,000	0,439	0,167
Dummy divorcerate	0,233	0,000	0,183	0,433
Dummy Houseprice	-0,227	0,007	-0,405	0,230
Dummy Refinancing	0,393	0,308	1,239	0,097
Propertytype PUD	-0,307	0,000	0,593	0,000
Propertytype CS	-0,107	0,002	0,391	0,002
Propertytype MH	0,136	0,000	0,044	0,749
Propertytype SFH	-0,039	0,339	-0,493	0,003
Purpose No Cash	-0,148	0,004	-0,295	0,114
Purpose Purchase	-0,220	0,000	0,216	0,111
Region 2	-0,456	0,315	-17,015	0,996
Region 3	0,067	0,247	-0,210	0,307
Region 4	-0,817	0,000	0,172	0,648
Region 5	-0,009	0,863	-0,170	0,308
Loan Age 2-3Y	0,550	0,000	0,884	0,000
Loan Age 3-4Y	0,365	0,000	1,170	0,000
Loan Age 4-6Y	0,201	0,000	1,077	0,000
Loan Age 6-8Y	0,092	0,088	0,940	0,000
Loan Age 8-10Y	0,026	0,686	0,442	0,078
Loan Age 10-15Y	-0,423	0,000	0,010	0,976

Table 37: Coefficient estimates and p-values Markov3 model before variable selection.

Part.Prep	Part.Prep		Prep.		Delinq.	
	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-6,513	0,000	-16,959	0,797	8,490	0,000
FICO score	0,006	0,000	0,003	0,032	-0,018	0,000
Firsthome	-0,153	0,293	-0,359	0,129	-0,014	0,969
Mortgage insurance	-0,013	0,010	0,003	0,715	0,002	0,861
DTI	-0,002	0,529	-0,003	0,590	-0,003	0,753
Loansize	-0,520	0,000	0,337	0,024	-0,310	0,246
LTV	-0,002	0,476	0,005	0,381	-0,003	0,720
Penalty	-13,338	0,985	2,011	0,078	-11,432	0,987
Unemployment	0,119	0,000	-0,144	0,001	0,086	0,145
Divorcerate	-0,069	0,219	-0,046	0,597	0,052	0,666
Houseprice	0,001	0,751	-0,005	0,102	-0,001	0,801
Refinancing	0,113	0,027	-0,319	0,000	-0,006	0,960
Dummy FICO	3,667	0,003	2,148	0,133	-11,236	0,000
Dummy Firsthome	-0,545	0,000	-0,037	0,845	-0,122	0,672
Dummy Insurance	0,522	0,122	0,894	0,029	-11,557	0,933
Dummy DTI	0,138	0,601	-0,212	0,682	0,310	0,662
Dummy Penalty	0,352	0,491	0,899	0,157	0,627	0,580
Dummy Unempl	1,179	0,000	-0,253	0,535	-0,765	0,507
Dummy divorcerate	-0,139	0,581	-0,070	0,852	-0,398	0,504
Dummy Houseprice	0,469	0,101	-1,106	0,051	-0,380	0,666
Dummy Refinancing	-14,080	0,978	2,014	0,017	-10,181	0,983
Propertytype PUD	-0,506	0,000	-0,265	0,153	-0,105	0,727
Propertytype CS	-0,151	0,164	-0,290	0,120	0,094	0,745
Propertytype MH	-0,082	0,527	0,307	0,156	-0,125	0,683
Propertytype SFH	-0,294	0,046	0,200	0,376	-0,511	0,145
Purpose No Cash	0,186	0,257	0,234	0,413	-0,158	0,659
Purpose Purchase	0,138	0,304	-0,058	0,801	-0,685	0,055
Region 2	-13,646	0,985	-12,985	0,986	-10,803	0,988
Region 3	0,418	0,075	-0,173	0,615	0,731	0,212
Region 4	-0,420	0,514	-0,673	0,525	0,754	0,410
Region 5	0,609	0,003	0,081	0,772	0,283	0,592
Loan Age 2-3Y	-0,651	0,000	12,397	0,851	0,194	0,646
Loan Age 3-4Y	-0,338	0,045	12,417	0,851	-0,164	0,727
Loan Age 4-6Y	-0,187	0,241	12,494	0,850	0,107	0,805
Loan Age 6-8Y	-0,120	0,499	12,630	0,848	-0,094	0,841
Loan Age 8-10Y	0,210	0,274	12,107	0,855	-0,998	0,125
Loan Age 10-15Y	0,582	0,005	12,528	0,850	0,402	0,476

Table 38: Coefficient estimates and p-values Markov5(1) model before variable selection.

Delinq.	Pa.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-2,441	0,047	-1,719	0,565	-1,049	0,087	-8,222	0,234
FICO score	0,000	0,894	0,005	0,102	-0,003	0,000	-0,003	0,601
Firsthome	0,366	0,090	-0,593	0,452	-0,206	0,085	-3,144	0,154
Mortgage ins	-0,007	0,374	-0,012	0,551	0,000	0,984	-0,016	0,701
DTI	-0,008	0,139	-0,007	0,639	0,016	0,000	0,048	0,110
Loansize	0,136	0,323	0,322	0,387	-0,168	0,020	0,400	0,581
LTV	-0,004	0,464	-0,021	0,082	0,007	0,043	0,031	0,485
Penalty	-10,922	0,979	-9,925	0,981	-12,112	0,976	-2,677	0,995
Unemployment	0,029	0,414	-0,484	0,000	0,063	0,000	0,273	0,117
Divorcerate	0,019	0,826	-0,006	0,978	-0,058	0,148	0,101	0,787
Houseprice	0,000	0,918	-0,003	0,587	0,002	0,084	0,006	0,717
Refinancing	-0,039	0,590	-0,299	0,130	-0,184	0,000	0,116	0,783
Dummy FICO	-0,075	0,956	-4,189	0,935	-2,022	0,006	-9,135	0,849
Dummy Firsthome	0,028	0,870	-0,082	0,849	0,033	0,686	-4,980	0,293
Dummy Insurance	0,171	0,737	-0,954	0,410	0,188	0,492	-3,321	0,850
Dummy DTI	-0,428	0,301	-0,346	0,775	0,600	0,006	-2,824	0,822
Dummy Penalty	-0,899	0,389	-7,185	0,798	-0,579	0,123	-3,876	0,869
Dummy Unempl	0,126	0,737	-4,491	0,001	-0,225	0,307	5,123	0,531
Dummy divorcerate	0,050	0,896	0,994	0,340	-0,201	0,270	1,648	0,384
Dummy Houseprice	0,702	0,145	-0,321	0,794	0,728	0,003	-3,757	0,649
Dummy Refinancing	0,266	0,746	-5,377	0,879	1,959	0,000	4,286	0,116
Propertytype PUD	-0,225	0,212	-0,411	0,369	0,277	0,002	-0,614	0,474
Propertytype CS	-0,114	0,530	-0,033	0,942	0,327	0,000	-1,013	0,281
Propertytype MH	-0,068	0,727	-0,247	0,596	0,054	0,594	1,044	0,246
Propertytype SFH	-0,014	0,946	-1,006	0,075	-0,020	0,856	-5,216	0,318
Purpose No Cash	-0,403	0,116	-1,496	0,160	-0,004	0,971	-5,198	0,427
Purpose Purchase	-0,508	0,011	-1,574	0,003	0,024	0,806	-0,822	0,492
Region 3	-0,260	0,466	-0,689	0,314	0,330	0,057	-1,362	0,301
Region 4	0,622	0,215	0,276	0,822	-0,058	0,835	-6,005	0,697
Region 5	0,076	0,795	-0,855	0,146	0,215	0,158	-1,548	0,068
Loan Age 2-3Y	0,535	0,036	0,620	0,192	0,454	0,000	-0,714	0,446
Loan Age 3-4Y	0,677	0,008	0,698	0,160	0,502	0,000	-5,659	0,238
Loan Age 4-6Y	0,933	0,000	-0,198	0,735	0,589	0,000	-0,433	0,636
Loan Age 6-8Y	1,108	0,000	-0,689	0,430	0,353	0,005	0,233	0,819
Loan Age 8-10Y	1,185	0,000	-0,425	0,706	0,828	0,000	-4,448	0,599
Loan Age 10-15Y	1,262	0,001	-6,381	0,578	0,057	0,784	-3,545	0,720

Table 39: Coefficient estimates and p-values Markov5(3) model before variable selection.

Contr.Paym.	Pa.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-6,266	0,000	-5,327	0,000	3,204	0,000	-4,157	0,000
FICO score	-0,001	0,000	0,001	0,000	-0,015	0,000	-0,012	0,000
Firsthome	0,074	0,049	-0,089	0,047	-0,111	0,094	0,005	0,977
Mortgage insurance	0,007	0,000	0,000	0,857	0,002	0,260	-0,002	0,706
DTI	-0,001	0,204	-0,007	0,000	0,014	0,000	0,029	0,000
Loansize	0,102	0,000	0,476	0,000	-0,010	0,824	0,399	0,000
LTV	-0,002	0,008	-0,003	0,001	0,010	0,000	0,030	0,000
Penalty	-0,230	0,518	-1,034	0,146	-1,018	0,311	0,918	0,364
Unemployment	0,245	0,000	-0,087	0,000	0,043	0,000	0,082	0,001
Divorcerate	0,019	0,200	0,103	0,000	-0,007	0,740	-0,021	0,690
Houseprice	0,001	0,007	-0,001	0,229	0,002	0,000	0,001	0,761
Refinancing	0,099	0,000	-0,706	0,000	-0,294	0,000	-0,530	0,000
Dummy FICO	0,055	0,835	1,160	0,000	-10,068	0,000	-24,828	0,992
Dummy Firsthome	0,231	0,000	-0,041	0,229	-0,106	0,036	-0,315	0,017
Dummy Insurance	0,348	0,001	0,697	0,000	0,480	0,002	0,354	0,490
Dummy DTI	-0,228	0,007	-0,392	0,000	0,815	0,000	1,449	0,000
Dummy Penalty	-0,104	0,574	0,078	0,648	0,235	0,290	0,717	0,125
Dummy Unempl	2,680	0,000	-0,148	0,085	0,400	0,002	0,522	0,106
Dummy divorcerate	0,081	0,236	0,211	0,001	-0,102	0,311	0,163	0,491
Dummy Houseprice	0,757	0,000	-0,203	0,019	0,185	0,150	-0,369	0,279
Dummy Refinancing	-1,285	0,000	-0,131	0,796	0,871	0,005	0,518	0,615
Propertytype PUD	-0,058	0,057	-0,301	0,000	0,169	0,001	0,627	0,000
Propertytype CS	-0,149	0,000	-0,094	0,008	0,187	0,000	0,433	0,001
Propertytype MH	0,116	0,001	0,147	0,000	-0,074	0,219	0,024	0,863
Propertytype SFH	0,442	0,000	-0,029	0,484	0,036	0,570	-0,453	0,006
Purpose No Cash	0,279	0,000	-0,130	0,014	-0,112	0,145	-0,262	0,162
Purpose Purchase	0,551	0,000	-0,192	0,000	0,284	0,000	0,255	0,063
Region 2	-0,441	0,199	-0,429	0,345	-18,354	0,996	-17,083	0,996
Region 3	0,106	0,038	0,086	0,146	0,207	0,033	-0,134	0,525
Region 4	-0,205	0,120	-0,816	0,000	0,344	0,036	0,258	0,497
Region 5	0,056	0,197	-0,003	0,955	0,227	0,006	-0,100	0,564
Loan Age 2-3Y	0,986	0,000	0,539	0,000	0,229	0,000	0,977	0,000
Loan Age 3-4Y	1,119	0,000	0,362	0,000	0,332	0,000	1,295	0,000
Loan Age 4-6Y	1,248	0,000	0,186	0,000	0,264	0,000	1,173	0,000
Loan Age 6-8Y	1,327	0,000	0,074	0,183	0,161	0,026	1,026	0,000
Loan Age 8-10Y	1,567	0,000	0,033	0,624	-0,255	0,012	0,542	0,033
Loan Age 10-15Y	2,146	0,000	-0,445	0,000	-0,232	0,067	0,129	0,700

Table 40: Coefficient estimates and p-values Markov5(5) model before variable selection.

D Model estimates Hausman test

	Prepayment	
	Coef.	P-value
C	-12,456	0,000
FICO score	-0,002	0,021
Firsthome	-0,434	0,008
Mortgage insurance	-0,020	0,001
DTI	-0,002	0,580
Loansize	-0,168	0,071
LTV	0,015	0,000
Penalty	-1,308	0,741
Unemployment	0,130	0,007
Divorcerate	-0,229	0,095
Houseprice	0,024	0,000
Refinancing	-1,435	0,000
Dummy Loan ID	20,253	0,000
Dummy FICO	-2,087	0,201
Dummy Firsthome	-1,092	0,000
Dummy Mortgage Insurance	2,740	0,000
Dummy DTI	0,247	0,569
Dummy Penalty	1,873	0,004
Dummy Unempl	-3,981	0,000
Dummy divorcerate	-0,827	0,187
Dummy Houseprice	-2,755	0,000
Dummy Refinancing	2,635	0,000
Propertytype Planned Unit Development	-0,060	0,678
Propertytype Cooperative Share	0,362	0,008
Propertytype Manufactured Housing	-0,407	0,001
Propertytype Single Family Home	-4,739	0,000
Purpose No Cash-out Refinance	-2,112	0,000
Purpose Purchase	-2,299	0,000
Region 2	-2,272	0,271
Region 3	-0,224	0,344
Region 4	-0,323	0,536
Region 5	-0,011	0,955
Loan Age 2-3Y	0,363	0,025
Loan Age 3-4Y	-0,455	0,010
Loan Age 4-6Y	-1,202	0,000
Loan Age 6-8Y	-2,316	0,000
Loan Age 8-10Y	-2,708	0,000
Loan Age 10-15Y	-3,432	0,000

Table 41: Coefficient estimates and p-values Hausman test three state MNL model (defaults removed).

	Pa.Prep		**Prep.**		**Delinq.**	
	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-4,717	0,000	-18,784	0,000	3,357	0,000
FICO score	-0,001	0,000	-0,001	0,365	-0,016	0,000
Firsthome	0,023	0,537	-0,205	0,187	-0,160	0,003
Mortgage insurance	0,003	0,015	-0,017	0,003	0,000	0,927
DTI	0,000	0,884	-0,006	0,148	0,020	0,000
Loansize	0,059	0,012	-0,065	0,471	-0,093	0,006
LTV	-0,001	0,512	0,010	0,007	0,014	0,000
Penalty	-0,590	0,128	-1,196	0,673	-1,430	0,155
Unemployment	0,022	0,000	0,178	0,000	0,059	0,000
Divorcerate	0,063	0,000	-0,230	0,035	-0,041	0,017
Houseprice	-0,001	0,165	0,019	0,000	0,004	0,000
Refinancing	-0,110	0,000	-1,157	0,000	-0,371	0,000
Dummy Loan ID	5,032	0,000	22,328	0,000	4,012	0,000
Dummy FICO	-0,449	0,084	-0,919	0,470	-10,765	0,000
Dummy Firsthome	-0,017	0,557	-0,900	0,000	-0,116	0,003
Dummy Insurance	0,246	0,010	2,464	0,001	0,531	0,000
Dummy DTI	-0,078	0,335	-0,143	0,733	0,999	0,000
Dummy Penalty	-0,005	0,976	1,658	0,078	-0,040	0,828
Dummy Unempl	0,022	0,727	-2,598	0,000	-0,125	0,226
Dummy divorcerate	0,303	0,000	-0,845	0,085	-0,269	0,001
Dummy Houseprice	0,180	0,016	-0,276	0,623	0,456	0,000
Dummy Refinancing	-0,114	0,667	2,210	0,001	1,857	0,000
Propertytype PUD	-0,041	0,180	-0,036	0,795	0,276	0,000
Propertytype CS	-0,048	0,108	0,269	0,044	0,332	0,000
Propertytype MH	0,041	0,230	-0,265	0,046	-0,070	0,133
Propertytype SFH	0,159	0,000	-2,978	0,000	-0,035	0,483
Purpose No Cash	0,068	0,124	-1,606	0,000	-0,092	0,111
Purpose Purchase	0,030	0,386	-1,714	0,000	0,258	0,000
Region 2	-0,427	0,241	-1,293	0,395	-10,704	0,864
Region 3	0,074	0,160	0,099	0,663	0,329	0,000
Region 4	-0,073	0,570	0,082	0,875	0,287	0,028
Region 5	0,125	0,005	0,250	0,201	0,282	0,000
Loan Age 2-3Y	1,172	0,000	0,885	0,000	0,540	0,000
Loan Age 3-4Y	1,317	0,000	0,573	0,001	0,633	0,000
Loan Age 4-6Y	1,452	0,000	0,068	0,684	0,616	0,000
Loan Age 6-8Y	1,637	0,000	-0,189	0,364	0,421	0,000
Loan Age 8-10Y	1,734	0,000	-0,544	0,018	0,181	0,014
Loan Age 10-15Y	2,049	0,000	-0,885	0,000	-0,083	0,381

Table 42: Coefficient estimates and p-values Hausman test five state MNL model (defaults removed).

E Coefficient estimates and p-values categorical variables

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
Dummy Loan ID	20,229	0,000	17,765	0,000
Dummy Firsthome	-1,049	0,000	-1,516	0,000
Dummy Mortgage Insurance	2,717	0,000	1,822	0,008
Dummy Penalty	1,735	0,006	1,924	0,006
Dummy Unempl	-4,572	0,000	-3,884	0,000
Dummy Houseprice	-2,835	0,000	-3,375	0,000
Dummy Refinancing	2,948	0,000	4,284	0,000
Propertytype Planned Unit Development	-0,039	0,772	1,205	0,000
Propertytype Cooperative Share	0,324	0,013	1,012	0,000
Propertytype Manufactured Housing	-0,358	0,004	-0,236	0,166
Propertytype Single Family Home	-4,396	0,000	-4,561	0,000
Purpose No Cash-out Refinance	-1,991	0,000	-1,871	0,000
Purpose Purchase	-2,108	0,000	-1,564	0,000
Region 2	-2,108	0,276	-5,363	0,238
Region 3	-0,198	0,379	-0,544	0,061
Region 4	-0,534	0,300	0,762	0,193
Region 5	0,022	0,907	-0,152	0,527
Loan Age 2-3Y	0,326	0,036	0,694	0,002
Loan Age 3-4Y	-0,522	0,002	0,401	0,086
Loan Age 4-6Y	-1,194	0,000	-0,187	0,415
Loan Age 6-8Y	-2,249	0,000	-1,202	0,000
Loan Age 8-10Y	-2,606	0,000	-1,922	0,000
Loan Age 10-15Y	-3,290	0,000	-2,798	0,000

Table 43: Coefficient estimates and p-values categorical variables MNL3 model.

	Pa.Prep		**Prep.**		**Delinq.**		**Default**	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
Dummy LoanID	5,013	0,000	19,403	0,000	3,587	0,000	17,244	0,000
Dummy Firsthome	-0,025	0,365	-0,849	0,000	-0,098	0,009	-1,047	0,000
Dummy Insurance	0,227	0,016	2,407	0,000	0,572	0,000	1,793	0,031
Dummy Penalty	-0,046	0,783	1,719	0,034	0,138	0,424	1,910	0,035
Dummy Unempl	0,039	0,527	-2,549	0,000	-0,019	0,855	-0,840	0,052
Dummy divorcerate	0,343	0,000	-1,024	0,021	-0,238	0,001	-1,122	0,024
Propertytype PUD	-0,109	0,681	2,302	0,001	2,432	0,000	4,145	0,000
Propertytype CS	-0,030	0,225	0,286	0,006	0,227	0,000	0,376	0,010
Propertytype MH	0,027	0,419	-0,171	0,166	0,010	0,820	-0,209	0,241
Propertytype SFH	0,159	0,000	-2,921	0,000	0,009	0,855	-3,304	0,000
Purpose No Cash	0,062	0,150	-1,562	0,000	0,053	0,338	-1,709	0,000
Purpose Purchase	0,008	0,824	-1,610	0,000	0,245	0,000	-1,117	0,000
Region 2	-0,439	0,229	-1,321	0,388	-7,951	0,529	-6,588	0,543
Region 3	0,085	0,105	0,104	0,633	0,231	0,002	-0,206	0,486
Region 4	-0,064	0,620	-0,090	0,861	0,763	0,000	1,310	0,031
Region 5	0,130	0,003	0,249	0,176	0,398	0,000	0,263	0,285
Loan Age 2-3Y	1,177	0,000	0,878	0,000	0,528	0,000	1,243	0,000
Loan Age 3-4Y	1,323	0,000	0,533	0,001	0,620	0,000	1,338	0,000
Loan Age 4-6Y	1,461	0,000	0,104	0,519	0,626	0,000	0,984	0,000
Loan Age 6-8Y	1,645	0,000	-0,212	0,289	0,485	0,000	0,625	0,026
Loan Age 8-10Y	1,741	0,000	-0,541	0,014	0,247	0,001	-0,080	0,808
Loan Age 10-15Y	2,063	0,000	-0,841	0,000	0,019	0,835	-0,623	0,118

Table 44: Coefficient estimates and p-values categorical variables MNL5 model.

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
Dummy FICO	1,319	0,000	-24,687	0,993
Dummy Mortgage Insurance	0,711	0,000	0,243	0,633
Dummy DTI	-0,380	0,000	1,498	0,000
Dummy Unempl	-0,267	0,001	0,487	0,110
Dummy divorcerate	0,213	0,001	0,185	0,425
Dummy Houseprice	-0,157	0,011	-0,524	0,025
Propertytype Planned Unit Development	-0,316	0,000	0,550	0,000
Propertytype Cooperative Share	-0,119	0,000	0,313	0,011
Propertytype Manufactured Housing	0,137	0,000	0,054	0,693
Propertytype Single Family Home	-0,040	0,323	-0,481	0,003
Purpose No Cash-out Refinance	-0,146	0,005	-0,286	0,124
Purpose Purchase	-0,220	0,000	0,221	0,101
Region 2	-0,459	0,312	-17,317	0,996
Region 3	0,067	0,245	-0,200	0,328
Region 4	-0,814	0,000	0,154	0,683
Region 5	-0,008	0,867	-0,147	0,379
Loan Age 2-3Y	0,543	0,000	0,900	0,000
Loan Age 3-4Y	0,358	0,000	1,200	0,000
Loan Age 4-6Y	0,193	0,000	1,117	0,000
Loan Age 6-8Y	0,083	0,111	0,982	0,000
Loan Age 8-10Y	0,016	0,800	0,474	0,057
Loan Age 10-15Y	-0,435	0,000	0,024	0,942

Table 45: Coefficient estimates and p-values categorical variables Markov3 model.

Part.Prep	Part.Prep		Prep.		Delinq.	
	Coef.	P-value	Coef.	P-value	Coef.	P-value
Dummy Insurance	0,179	0,582	1,107	0,004	-11,505	0,942
Dummy Refinancing	-13,642	0,979	1,694	0,036	-11,426	0,983
Propertytype PUD	-0,502	0,000	-0,311	0,058	0,044	0,861
Propertytype CS	-0,151	0,111	-0,300	0,072	0,152	0,531
Propertytype MH	-0,095	0,444	0,354	0,093	-0,068	0,814
Propertytype SFH	-0,340	0,015	0,260	0,233	-0,476	0,156
Purpose No Cash	0,076	0,624	0,223	0,428	0,080	0,814
Purpose Purchase	-0,076	0,534	0,075	0,722	-0,766	0,020
Region 2	-13,749	0,986	-12,992	0,987	-11,847	0,988
Region 3	0,417	0,071	-0,140	0,683	0,702	0,221
Region 4	-0,355	0,570	-0,602	0,567	1,109	0,213
Region 5	0,560	0,005	0,113	0,685	0,416	0,423
Loan Age 2-3Y	-0,625	0,000	12,485	0,857	0,197	0,634
Loan Age 3-4Y	-0,244	0,140	12,459	0,858	-0,239	0,601
Loan Age 4-6Y	-0,068	0,663	12,499	0,857	0,050	0,905
Loan Age 6-8Y	0,003	0,985	12,604	0,856	-0,022	0,962
Loan Age 8-10Y	0,331	0,077	12,093	0,862	-1,007	0,114
Loan Age 10-15Y	0,784	0,000	12,484	0,857	0,152	0,778

Table 46: Coefficient estimates and p-values categorical variables Markov5(1) model.

Delinq.	Pa.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
Dummy Unempl	0,533	0,095	-4,051	0,001	0,182	0,299	1,300	0,362
Dummy divorcerate	0,230	0,161	0,985	0,021	0,093	0,260	0,447	0,615
Propertytype PUD	-0,268	0,095	-0,249	0,564	0,299	0,000	-0,502	0,540
Propertytype CS	-0,148	0,360	-0,003	0,994	0,338	0,000	-0,952	0,280
Propertytype MH	-0,113	0,541	-0,227	0,611	0,059	0,541	0,673	0,392
Propertytype SFH	-0,019	0,922	-0,896	0,099	-0,027	0,798	-8,178	0,686
Purpose No Cash	-0,465	0,062	-1,539	0,143	0,032	0,784	-7,929	0,751
Purpose Purchase	-0,642	0,001	-1,490	0,003	-0,039	0,675	-0,506	0,591
Region 3	-0,124	0,710	-0,518	0,433	0,117	0,470	-1,772	0,146
Region 4	0,502	0,293	0,133	0,910	-0,155	0,568	-9,436	0,877
Region 5	0,106	0,703	-0,799	0,167	0,034	0,810	-1,801	0,020
Loan Age 2-3Y	0,496	0,051	0,549	0,244	0,502	0,000	-0,602	0,519
Loan Age 3-4Y	0,629	0,013	0,528	0,275	0,555	0,000	-8,205	0,661
Loan Age 4-6Y	0,903	0,000	-0,315	0,578	0,696	0,000	-0,216	0,799
Loan Age 6-8Y	1,099	0,000	-0,780	0,351	0,486	0,000	0,253	0,785
Loan Age 8-10Y	1,174	0,000	-0,530	0,633	0,963	0,000	-7,509	0,820
Loan Age 10-15Y	1,362	0,000	-9,009	0,834	0,300	0,128	-7,173	0,863

Table 47: Coefficient estimates and p-values categorical variables Markov5(3) model.

Contr.Pay	Pa.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
Dummy FICO	-0,055	0,834	1,163	0,000	-10,141	0,000	-25,310	0,993
Dummy Insurance	0,077	0,441	0,723	0,000	0,331	0,030	0,270	0,595
Dummy DTI	-0,180	0,031	-0,388	0,000	0,866	0,000	1,533	0,000
Dummy divorcerate	0,301	0,000	0,185	0,004	-0,001	0,991	0,180	0,440
Dummy Houseprice	1,316	0,000	-0,189	0,001	-0,018	0,823	-0,264	0,166
Propertytype PUD	-0,012	0,689	-0,311	0,000	0,154	0,002	0,585	0,000
Propertytype CS	-0,044	0,113	-0,106	0,002	0,153	0,002	0,358	0,004
Propertytype MH	0,054	0,103	0,151	0,000	-0,073	0,217	0,041	0,768
Propertytype SFH	0,213	0,000	-0,023	0,575	0,032	0,614	-0,453	0,005
Purpose No Cash	0,101	0,019	-0,123	0,020	-0,118	0,121	-0,264	0,157
Purpose Purchase	0,154	0,000	-0,177	0,000	0,251	0,000	0,235	0,083
Region 2	-0,323	0,344	-0,435	0,339	-18,745	0,997	-17,519	0,997
Region 3	0,081	0,108	0,087	0,142	0,205	0,035	-0,126	0,550
Region 4	-0,116	0,368	-0,813	0,000	0,373	0,023	0,248	0,514
Region 5	0,067	0,115	-0,003	0,950	0,234	0,004	-0,077	0,657
Loan Age 2-3Y	1,031	0,000	0,530	0,000	0,270	0,000	1,000	0,000
Loan Age 3-4Y	1,149	0,000	0,350	0,000	0,389	0,000	1,334	0,000
Loan Age 4-6Y	1,287	0,000	0,174	0,000	0,334	0,000	1,221	0,000
Loan Age 6-8Y	1,411	0,000	0,065	0,228	0,226	0,001	1,070	0,000
Loan Age 8-10Y	1,547	0,000	0,021	0,752	-0,205	0,041	0,570	0,024
Loan Age 10-15Y	1,889	0,000	-0,455	0,000	-0,202	0,107	0,100	0,764

Table 48: Coefficient estimates and p-values categorical variables Markov5(5) model.

	Prior	**Crisis**			**Post**	**Crisis**		
	Prep		Default		Prep		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
Dummy Loan ID	20,965	0,000	17,778	0,000	18,743	0,000	16,852	0,000
Dummy Firsthome	0,440	0,330	-0,268	0,554	-0,595	0,000	-0,837	0,000
Dummy Insurance	0,499	0,651	0,134	0,907	-1,909	0,874	0,722	0,956
Dummy Penalty	0,209	0,919	0,380	0,852	1,974	0,045	2,332	0,027
Dummy Unempl	0,761	0,641	-0,807	0,605	-5,133	0,000	-4,333	0,000
Dummy Houseprice	0,066	0,968	2,539	0,108	-3,167	0,000	-4,191	0,000
Dummy Refinancing	0,718	0,842	-1,778	0,738	-2,108	0,716	5,376	0,000
Propertytype PUD	-0,348	0,447	0,860	0,059	-0,379	0,023	0,887	0,000
Propertytype CS	-0,179	0,689	0,732	0,101	0,080	0,618	0,583	0,010
Propertytype MH	-0,119	0,818	0,468	0,368	0,061	0,705	-0,304	0,180
Propertytype SFH	0,350	0,527	0,223	0,693	-4,273	0,000	-4,530	0,000
Purpose No Cash	-0,041	0,952	0,428	0,531	-1,427	0,000	-1,690	0,000
Purpose Purchase	-0,245	0,635	0,597	0,248	-1,922	0,000	-1,717	0,000
Region 2	1,621	0,787	0,175	0,981	-2,108	0,312	-5,894	0,287
Region 3	1,207	0,179	0,913	0,313	-0,101	0,695	-0,582	0,096
Region 4	0,566	0,771	1,456	0,439	-0,748	0,251	0,882	0,201
Region 5	1,102	0,160	1,001	0,204	-0,198	0,358	-0,485	0,093
Loan Age 2-3Y	-0,080	0,860	0,225	0,620	1,000	0,000	1,871	0,000
Loan Age 3-4Y	-0,286	0,578	0,311	0,544	0,051	0,834	1,663	0,000
Loan Age 4-6Y	-0,374	0,460	0,372	0,459	-0,359	0,129	1,105	0,006
Loan Age 6-8Y	-0,425	0,573	0,431	0,562	-1,190	0,000	0,175	0,688
Loan Age 8-10Y	-0,263	0,872	0,109	0,946	-1,323	0,000	-0,363	0,429
Loan Age 10-15Y	-0,918	0,737	0,662	0,803	-2,001	0,000	-1,386	0,009

Table 49: Coefficient estimates and p-values categorical variables MNL3 model with structural breaks.

F Models including put option

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
C	-13,489	0,000	-17,548	0,000
Firsthome	-0,422	0,005	-0,312	0,140
Mortgage insurance	-0,013	0,017	-0,006	0,425
Loansize	-0,455	0,000	-0,351	0,005
LTV	0,020	0,000	0,057	0,000
Houseprice	0,017	0,000	0,016	0,000
Refinancing	-0,002	0,985	0,067	0,733
Dummy Loan ID	20,837	0,000	18,288	0,000
Dummy Firsthome	-1,297	0,000	-1,759	0,000
Dummy Mortgage Insurance	2,452	0,000	1,423	0,035
Dummy Penalty	2,226	0,001	2,565	0,001
Dummy Unempl	-4,681	0,000	-4,134	0,000
Dummy Houseprice	-4,211	0,000	-5,100	0,000
Dummy Refinancing	0,999	0,142	1,966	0,037
Propertytype Planned Unit Development	0,274	0,040	1,540	0,000
Propertytype Cooperative Share	0,370	0,004	1,030	0,000
Propertytype Manufactured Housing	-0,549	0,000	-0,439	0,010
Propertytype Single Family Home	-4,730	0,000	-4,910	0,000
Purpose No Cash-out Refinance	-2,139	0,000	-2,012	0,000
Purpose Purchase	-2,344	0,000	-1,814	0,000
Region 2	-1,844	0,402	-5,343	0,309
Region 3	-0,297	0,174	-0,679	0,019
Region 4	-0,417	0,432	0,873	0,154
Region 5	0,097	0,596	-0,106	0,658
Loan Age 2-3Y	0,171	0,255	0,602	0,009
Loan Age 3-4Y	0,044	0,788	1,084	0,000
Loan Age 4-6Y	-0,272	0,082	0,913	0,000
Loan Age 6-8Y	0,058	0,752	1,344	0,000
Loan Age 8-10Y	-0,340	0,094	0,593	0,048
Loan Age 10-15Y	-1,071	0,000	-0,221	0,550

Table 50: Coefficient estimates and p-values MNL3 model with put option.

	Pa.Prep		**Prep.**		**Delinq.**		**Default**	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-5,562	0,000	-17,554	0,000	-7,246	0,000	-22,376	0,000
Mortgage insurance	0,004	0,000	-0,011	0,036	0,009	0,000	-0,008	0,275
LTV	0,001	0,453	0,011	0,002	0,017	0,000	0,040	0,000
Unemployment	0,033	0,000	0,212	0,000	0,062	0,000	0,398	0,000
Divorcerate	0,067	0,000	-0,219	0,027	-0,031	0,057	-0,310	0,006
Houseprice	-0,002	0,000	0,022	0,000	0,001	0,000	0,027	0,000
Refinancing	0,043	0,464	0,043	0,680	0,122	0,000	0,114	0,479
Dummy LoanID	4,971	0,000	19,898	0,000	3,313	0,000	17,602	0,000
Dummy Firsthome	-0,023	0,412	-0,950	0,000	-0,069	0,064	-1,139	0,000
Dummy Insurance	0,196	0,038	2,319	0,001	0,348	0,004	1,658	0,047
Dummy Penalty	0,001	0,996	2,130	0,012	0,428	0,012	2,428	0,010
Dummy Unempl	0,079	0,199	-2,421	0,000	0,075	0,446	-0,809	0,056
Dummy divorcerate	0,345	0,000	-1,061	0,013	-0,249	0,001	-1,144	0,019
Propertytype PUD	-0,191	0,471	1,038	0,117	1,914	0,000	2,520	0,006
Propertytype CS	-0,053	0,030	0,123	0,227	0,111	0,000	0,173	0,233
Propertytype MH	0,020	0,550	-0,229	0,052	-0,035	0,436	-0,283	0,106
Propertytype SFH	0,153	0,000	-3,082	0,000	-0,024	0,626	-3,488	0,000
Purpose No Cash	0,067	0,121	-1,577	0,000	0,077	0,163	-1,725	0,000
Purpose Purchase	0,022	0,520	-1,672	0,000	0,313	0,000	-1,183	0,000
Region 2	-0,418	0,250	-0,755	0,621	-8,535	0,627	-6,466	0,640
Region 3	0,081	0,123	-0,020	0,923	0,208	0,005	-0,362	0,207
Region 4	-0,040	0,757	0,091	0,861	0,886	0,000	1,507	0,014
Region 5	0,133	0,003	0,302	0,080	0,409	0,000	0,291	0,220
Loan Age 2-3Y	1,200	0,000	0,819	0,000	0,675	0,000	1,220	0,000
Loan Age 3-4Y	1,369	0,000	0,959	0,000	0,890	0,000	1,846	0,000
Loan Age 4-6Y	1,530	0,000	0,861	0,000	1,049	0,000	1,867	0,000
Loan Age 6-8Y	1,746	0,000	1,578	0,000	1,050	0,000	2,592	0,000
Loan Age 8-10Y	1,879	0,000	1,308	0,000	0,957	0,000	1,966	0,000
Loan Age 10-15Y	2,268	0,000	1,109	0,000	1,155	0,000	1,639	0,000

Table 51: Coefficient estimates and p-values MNL5 model with put option.

	Prepayment		**Default**	
	Coef.	P-value	Coef.	P-value
C	-4,914	0,000	-3,842	0,000
FICO score	0,000	0,567	-0,012	0,000
Firsthome	-0,081	0,061	0,053	0,734
DTI	-0,003	0,003	0,031	0,000
Loansize	0,309	0,000	0,306	0,001
LTV	-0,001	0,430	0,031	0,000
Unemployment	0,004	0,496	0,148	0,000
Divorcerate	0,078	0,000	-0,017	0,744
Refinancing	0,097	0,007	0,144	0,288
Dummy FICO	0,569	0,047	-25,347	0,992
Dummy Mortgage Insurance	0,451	0,000	0,047	0,927
Dummy DTI	-0,223	0,016	1,607	0,000
Dummy Unempl	0,271	0,001	0,807	0,007
Dummy divorcerate	0,298	0,000	0,288	0,234
Dummy Houseprice	-0,309	0,000	-0,503	0,032
Propertytype Planned Unit Development	-0,306	0,000	0,558	0,000
Propertytype Cooperative Share	-0,227	0,000	0,232	0,059
Propertytype Manufactured Housing	0,012	0,738	-0,012	0,931
Propertytype Single Family Home	-0,122	0,003	-0,520	0,001
Purpose No Cash-out Refinance	-0,187	0,000	-0,269	0,147
Purpose Purchase	-0,214	0,000	0,209	0,126
Region 2	-0,395	0,383	-17,180	0,996
Region 3	0,068	0,240	-0,173	0,399
Region 4	-0,719	0,001	0,208	0,581
Region 5	0,003	0,952	-0,123	0,461
Loan Age 2-3Y	0,667	0,000	1,016	0,000
Loan Age 3-4Y	0,611	0,000	1,421	0,000
Loan Age 4-6Y	0,618	0,000	1,457	0,000
Loan Age 6-8Y	0,701	0,000	1,414	0,000
Loan Age 8-10Y	0,812	0,000	1,007	0,000
Loan Age 10-15Y	0,876	0,000	0,950	0,003

Table 52: Coefficient estimates and p-values Markov3 model with put option.

Part.Prep	Part.Prep		Prep.		Delinq.	
	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	0,542	0,212	-18,061	0,777	-8,937	0,000
Loansize	-0,653	0,000	0,429	0,005	-0,113	0,660
Penalty	-13,887	0,988	1,960	0,082	-12,495	0,992
Unemployment	-0,017	0,341	-0,059	0,043	0,110	0,015
Refinancing	-25,440	0,000	20,099	0,000	36,802	0,000
Dummy Insurance	0,623	0,059	0,734	0,059	-12,284	0,948
Dummy Refinancing	-14,772	0,986	1,526	0,060	-12,686	0,990
Propertytype PUD	-0,555	0,000	-0,250	0,129	0,174	0,490
Propertytype CS	-0,239	0,012	-0,274	0,103	0,322	0,192
Propertytype MH	-0,131	0,297	0,368	0,080	0,042	0,886
Propertytype SFH	-0,331	0,018	0,231	0,287	-0,478	0,156
Purpose No Cash	0,116	0,459	0,213	0,448	0,035	0,918
Purpose Purchase	-0,106	0,385	0,080	0,703	-0,672	0,042
Region 2	-13,312	0,984	-12,753	0,986	-11,831	0,988
Region 3	0,362	0,118	-0,085	0,804	0,789	0,169
Region 4	-0,348	0,578	-0,535	0,610	1,172	0,188
Region 5	0,567	0,004	0,119	0,667	0,397	0,444
Loan Age 2-3Y	-0,638	0,000	12,395	0,846	0,212	0,611
Loan Age 3-4Y	-0,268	0,105	12,459	0,845	-0,184	0,688
Loan Age 4-6Y	-0,068	0,656	12,610	0,843	0,134	0,746
Loan Age 6-8Y	0,011	0,947	12,835	0,841	0,111	0,801
Loan Age 8-10Y	0,240	0,177	12,443	0,845	-0,753	0,230
Loan Age 10-15Y	0,739	0,000	12,941	0,839	0,281	0,572

Table 53: Coefficient estimates and p-values Markov5(1) model with put option.

Delinq.	Pa.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-1,715	0,117	-6,681	0,028	-1,497	0,005	-5,558	0,267
FICO score	0,000	0,721	0,007	0,025	-0,002	0,001	0,000	0,948
LTV	-0,004	0,357	-0,024	0,013	0,008	0,003	0,017	0,568
Unemployment	0,030	0,353	-0,316	0,004	0,068	0,000	0,223	0,125
Refinancing	-5,658	0,238	21,163	0,046	4,693	0,034	1,506	0,948
Dummy Unempl	0,536	0,090	-3,525	0,005	0,370	0,033	1,260	0,377
Dummy divorcerate	0,193	0,246	1,208	0,008	0,124	0,135	0,401	0,650
Propertytype PUD	-0,252	0,116	-0,187	0,664	0,327	0,000	-0,596	0,462
Propertytype CS	-0,143	0,376	-0,010	0,982	0,341	0,000	-0,955	0,278
Propertytype MH	-0,123	0,505	-0,247	0,579	0,028	0,769	0,665	0,395
Propertytype SFH	-0,018	0,924	-0,803	0,142	-0,030	0,775	-8,250	0,689
Purpose No Cash	-0,432	0,085	-1,624	0,123	0,021	0,860	-8,054	0,750
Purpose Purchase	-0,628	0,001	-1,442	0,005	-0,015	0,868	-0,595	0,518
Region 3	-0,136	0,683	-0,411	0,536	0,145	0,372	-1,736	0,154
Region 4	0,535	0,262	0,202	0,864	-0,135	0,620	-9,579	0,878
Region 5	0,108	0,698	-0,778	0,177	0,034	0,813	-1,781	0,019
Loan Age 2-3Y	0,499	0,049	0,583	0,216	0,529	0,000	-0,636	0,496
Loan Age 3-4Y	0,646	0,011	0,610	0,207	0,612	0,000	-8,325	0,661
Loan Age 4-6Y	0,921	0,000	-0,172	0,761	0,791	0,000	-0,321	0,702
Loan Age 6-8Y	1,135	0,000	-0,611	0,458	0,622	0,000	0,117	0,897
Loan Age 8-10Y	1,207	0,000	-0,311	0,777	1,139	0,000	-7,732	0,819
Loan Age 10-15Y	1,463	0,000	-8,639	0,844	0,580	0,002	-7,596	0,858

Table 54: Coefficient estimates and p-values Markov5(3) model with put option.

Contr.Pay	Pa.Prep		Prep.		Delinq.		Default	
	Coef.	P-value	Coef.	P-value	Coef.	P-value	Coef.	P-value
C	-5,397	0,000	-5,464	0,000	3,501	0,000	-3,948	0,000
FICO score	-0,001	0,000	0,001	0,000	-0,015	0,000	-0,011	0,000
Firsthome	0,051	0,167	-0,096	0,031	-0,098	0,141	-0,009	0,956
DTI	0,000	0,945	-0,004	0,001	0,014	0,000	0,032	0,000
Loansize	0,016	0,752	0,072	0,278	0,089	0,337	0,106	0,724
LTV	0,002	0,043	-0,002	0,027	0,011	0,000	0,029	0,000
Unemployment	0,096	0,000	-0,058	0,000	0,027	0,002	0,082	0,000
Divorcerate	0,049	0,000	0,099	0,000	0,012	0,543	-0,011	0,837
Refinancing	0,037	0,003	-0,660	0,000	-0,280	0,000	-0,510	0,000
Dummy FICO	-0,031	0,906	1,338	0,000	-10,146	0,000	-25,254	0,994
Dummy Insurance	0,049	0,621	0,589	0,000	0,333	0,028	0,161	0,751
Dummy DTI	-0,161	0,053	-0,271	0,005	0,860	0,000	1,658	0,000
Dummy divorcerate	0,314	0,000	0,232	0,000	-0,003	0,971	0,236	0,309
Dummy Houseprice	1,324	0,000	-0,138	0,014	-0,018	0,820	-0,230	0,224
Propertytype PUD	-0,009	0,746	-0,291	0,000	0,154	0,002	0,590	0,000
Propertytype CS	-0,038	0,170	-0,067	0,047	0,154	0,002	0,362	0,004
Propertytype MH	0,020	0,536	-0,024	0,517	-0,069	0,235	-0,109	0,419
Propertytype SFH	0,208	0,000	-0,066	0,114	0,032	0,616	-0,471	0,004
Purpose No Cash	0,064	0,128	-0,306	0,000	-0,113	0,131	-0,426	0,019
Purpose Purchase	0,128	0,000	-0,303	0,000	0,255	0,000	0,124	0,350
Region 2	-0,328	0,337	-0,508	0,264	-18,818	0,997	-17,635	0,997
Region 3	0,101	0,043	0,197	0,001	0,199	0,038	-0,010	0,961
Region 4	-0,143	0,267	-0,948	0,000	0,374	0,023	0,171	0,652
Region 5	0,073	0,087	0,035	0,495	0,232	0,005	-0,055	0,749
Loan Age 2-3Y	1,029	0,000	0,504	0,000	0,271	0,000	0,984	0,000
Loan Age 3-4Y	1,143	0,000	0,312	0,000	0,391	0,000	1,310	0,000
Loan Age 4-6Y	1,278	0,000	0,126	0,004	0,335	0,000	1,193	0,000
Loan Age 6-8Y	1,398	0,000	0,004	0,948	0,228	0,001	1,026	0,000
Loan Age 8-10Y	1,530	0,000	-0,063	0,335	-0,202	0,044	0,493	0,050
Loan Age 10-15Y	1,864	0,000	-0,585	0,000	-0,198	0,112	-0,008	0,982

Table 55: Coefficient estimates and p-values Markov5(5) model with put option.

G Quandt's Likelihood Ratio Test

An alternative to the CUSUM tests is Quandt's likelihood ratio test (QLR) which is a likelihood ratio test for testing $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$ when the break point t_b is unknown. The QLR test statistic is defined as the maximum Chow test statistic, $F_T(\lambda)$ defined by

$$F_T(\lambda) = \frac{(SSR_{1,T} - (SSR_{1,t_b} + SSR_{t_b+1,T})/k}{(SSR_{1,t_b} + SSR_{t_b+1,T})/(T-2k)} \sim \; F_{k,T-2k} \tag{G.1}$$

in which $SSR_{\tau,t} = \hat{\epsilon}'_{\tau,t}\hat{\epsilon}_{\tau,t}$ is the sum of squared residuals (SSR) from the model using observations $\tau, ...t$.

The maximum is taken over a range of break dates $t_0, ..., t_1$

$$supF = \max_{t_b \in [t_0,...,t_1]} F_T(\frac{t_b}{T}) = \max_{\lambda \in [\lambda_0,...,\lambda_1]} F_T(\lambda), \tag{G.2}$$

in which $\lambda_i = \ddot{}i/T$ are trimming parameters, $i = 0, 1$. If there is no knowledge of the break date the parameters can be set as $\lambda_0 = 0.15$ and $\lambda_1 = 0.85$ (Andrews, 1993). The trimming fraction is applied to ensure reliable/accurate parameter estimates before and after the break. Each individual $F_T(\lambda)$ follows a $\chi^2(k)$ distribution. Due to possible correlations between $F_T(\lambda_1)$ and $F_T(\lambda_2)$, the distribution of *supF* is complicated to derive analytically. (Andrews, 1993) derived the distribution of *supF* numerically and shows that it depends on both k and λ.

The location of *supF* coincides with the location where the residual variance is minimized. This location can be determined by estimating the model for each possible break date, $\lambda_0 T < t_b < \lambda_1 T$, and computing the variance of the residuals at each point in time, $\hat{\sigma}^2(\lambda) = \frac{1}{T}\sum_{t=1}^{T} \hat{\epsilon}_t^2(\lambda)$. An estimate of t_b can be obtained by minimizing this residual variance

$$\hat{t}_b = \min{(\hat{\sigma}_t^2(t_b))}. \tag{G.3}$$

References

Federal Housing Finance Agency. The size of the affordable mortgage market: 2015-2017 enterprise single-family housing goals. Technical report, Federal Housing Finance Agency, 2014.

P.D. Allison. Measures of fit for logistic regression. SAS Global Forum, Paper 1485, 2014.

D.W.K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 59(5826):817–858, 1993.

E.K. Baldvinsdottir and L. Palmborg. On constructing a market consistent economic scenario generator. Handelsbanken Liv, 2011.

M. Bissiri and R. Cogo. Modelling behavioral risk. Cassa Depositi e Prestiti S.p.a, 2014.

T. Bjork. *Arbitrage theory in continuous time.* Oxford University Press, Third Edition, 2009.

D. Brigo. *Interest rate models - Theory and practise. With smile, inflation and credit.* Springeer Finance, 2010.

W. Burns. Prepayment modelling challenges in the wake of the 2008 credit and mortgage crisis. *Interactive Data Fixed Income Analytics*, 0930:1–8, 2010.

A. Van Bussel. *Valuation and interest rate risk of mortgages in the Netherlands.* PhD thesis, Maastricht University, 1998.

E. Charlier and A. Van Bussel. Prepayment behavior of dutch mortgagors. *Center Discussion Paper, Tilburg: Econometrics*, 2001(64):1–33, 2001.

J.M. Clapp, G.M. Goldberg, J.P. Harding, and M. LaCour-Little. Movers and shuckers: Interdependent prepayment decisions. Center for Real Estate, University of Conneticut, 2000.

M. Consalvi and G. Scotto di Freca. Measuring prepayment risk: an application to unicredit family financing. *UniCredit Universities Working Paper Series*, (05):1–35, 2010.

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

Y. Deng, J.M. Clapp, and X. An. Unobserved heterogeneity in models of competing mortgage termination. Social Science Research Network, 2005.

K. Gerardi and C. Hudson. Did nonrecourse mortgages cause the mortgage crisis? *Real Estate Research Federal Reserve Bank of Atlanta*, 2010.

Y. Gonchanov. *An intensity based approach for valuation of mortgage contracts subject to prepayment risk.* PhD thesis, University of Illinois, Department of Mathematics, Statistics and Computer Science, 2002.

J. Green and J.B. Shoven. The effects of interest rates on mortgage prepayments. *National Bureau of Economic Research*, (1246):1–32, 1983.

S. Gudell. Q1 2015: Negative equity report: After three long years, the hard work begins now. *Zillow Real Estate Research*, 2015.

J.P.A.M. Jacobs, R.H. Koning, and E. Sterken. Modelling prepayment risk. University of Groningen, 2005.

A. Kolbe. *Valuation of mortgage products with stochastic prepayment-intensity models*. PhD thesis, Technical University of Munchen, Centre for Mathematics, 2008.

S. Perry, S. Robinson, and J. Rowland. A study of mortgage prepayment risk. *Housing Finance International*, pages 36–51, 2001.

M. Plomp. Economic scenario generator, 2013. KPMG.

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

G. Rodriguez. Non parametric estimation in survival models. Princeton, 2005.

H. Tanizaki. Asymptotically exact confidence intervals of cusum and cusumsq tests: A numerical derivation using simulation techniques. *Communications in Statistics - Simulation and Computation*, 24(4):1019–1036, 2007.

P. Vasconcelos. Modelling prepayment risk: Multinomial logit approach for assessing conditional prepayment rate, 2010. NIBC.

W. Vijverberg. Testing for iia with the hausman-mcfadden test. *IZA Discussion Paper*, (5826), 2011.