Erasmus University Rotterdam

Erasmus School of Economics

FEM21030 - Master Thesis Econometrics
(Quantitative Finance)

Prediction and Modelling of Mortgage Prepayment Risk in a Low Interest Rate Environment Using Time-Varying Parameters

Author
T.F. Wesseling
ERNA ID: 436683tw

Supervisor Deloitte
P.G.A. Abeling

Supervisor
Dr. R. Lange

Co-reader
A.A. Naghi

2nd August 2018

Abstract

In this paper we estimate the prepayment risk of American mortgages on the residential market. Since the housing prices increase, whereas the mortgage interest rates decrease, we believe that correlations and variable coefficients differ before and after the financial crisis of 2008. The currently used survival analysis, MNL model and Markov switching model fail to take the fluctuation of parameters over time into account, but show differences between distinct intervals. We use a time-varying Markov switching model with dynamic parameters and a generalized auto-regressive score to investigate the presence of time-dependency. Both the results of five state and two state model are not significant, meaning that we find no evidence of time-dependency. Overall we conclude that we are unable to improve upon the currently used MNL model in terms of mortgage prepayment estimation.

Keywords: Mortgages, Prepayment Risk, Regime Switching, Time-Varying Parameters, GAS model

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information Criterion	19, 27, 28, 40, 41, 51
ARM	Adjustable Rate Mortgage	3, 6, 7
BIC	Bayesian Information Criterion	40, 51
DTI	Debt-to-Income	13, 15, 26–28, 31, 42, 43, 50
Fannie Mae	Federal National Mortgage Association	6
FDFD	Federal Funds Rate	9
FED	Federal Resevere, Central Bank of the US	5, 17
FHFA	Federal Housing Finance Agency	7
Freddie Mac	Federal Home Loan Mortgage Coporation	6
FRM	Fixed Rate Mortgage	3, 6, 7
GAS	generalized autoregressive score	24, 36, 37, 39, 40, 44
Ginnie Mae	Government National Mortgage Association	6
LTV	Loan-to-Value	8, 13, 15, 27, 28, 31, 42
MAE	Mean Absolute Error	40, 51, 52
MASE	Mean Absolute Scaled Error	40, 52
MBS	Mortgage-Backed Security	1, 6
MNL	Multinomial Logit	1, 2, 4, 9, 10, 16, 18, 19, 24, 29, 30, 33, 41–44, 59–63
MSE	Mean Squared Error	40, 52
MSSE	Mean Squared Scaled Error	40, 52
TVP	Time-Varying Probabilities	24, 37, 39, 40

1. Introduction

1.1. Research Design

As of June 2017, the American mortgage market consists of 14.6 trillion dollars, of which 80.0% is residential (Reserve, 2017). This makes the US mortgage market one of the biggest global fixed income markets (Chernov et al., 2016). Buyers of these mortgages are commonly referred to as mortgagors, whereas the seller of a mortgage is called a mortgagee. Since a mortgage is simply a long-term loan with the house as a collateral, there are risks involved for both the mortgagor and the mortgagee. In this research the perspective of the mortgagee is taken, whose biggest risk is premature payment of the loan, also referred to as prepayment risk. After the financial crisis of 2008 the American mortgage market is rapidly changing. Housing prices go up whereas the savings rate remains low. This implies that most models, often based on historical data, might not be sufficiently capable of estimating the present dynamics of mortgage prepayment.

Models that are currently used to estimate mortgage prepayment are static and do not take time- dependency into account. Vasconcelos (2010) and Meis (2015) use a MNL model that explains well but assumes that all observations are independent and consecutive. Schwartz and Torous (1989) uses a survival analysis to capture the prepayment function in order to explain the stylized facts of a Mortgage-Backed Security (MBS). Meis (2015) also uses a static Markov Switching model, but concludes that this model is unable to capture auto-correlation between the prepayment rates. Since macroeconomic variables change over time and influence the behaviour of the mortgagor, above methods might lead to inaccurate estimates of the prepayment rates. Therefore, we use the time-varying probability model of Bazzi et al. (2017), who built a dynamic framework which drops the assumption of static parameters.

The main goal of this research is to look for time-dependency within residential mortgage prepayments in order to get better estimates of the prepayment rate and frequency. To capture time-dependency a GAS model is built where transition probabilities are time-varying. As benchmark the MNL model and a survival analysis are applied, since these are most frequently used by mortgagees to estimate prepayment risk. Because these models are hard to compare with the GAS model, we also created a static Markov Switching model. Mortgages are obtained from the Freddie Mac Database, which consists of dynamic monthly performance and static mortgage specific observations of fixed rate US mortgages originated between 2000 and 2016.

With the three-state and five-state MNL model we can compare the relative likelihood of being in any payment state compared to paying on schedule. This shows that an increase in FICO score or a decrease in interest rate, LTV and DTI leads to an increase of prepayments and might also lower the delinquency rate. Furthermore, sub-segmentation shows that most partial prepayments are no more than 10% of the outstanding debt. The loan age of the mortgage is used for the survival analysis and tells us that the termination rate decreases for a longer holding period. Moreover, half of the mortgages is terminated within five years. Loan Age, CLTV and the interest rate shorten the mortgage lifetime, whereas LTV and DTI have a positive effect on the expected lifetime of the mortgage. A static Markov Switching model is built to gain insight in the probability to move from one state to another. The five-state Markov models tell us that mortgagors tend to stay in the same payment state over time. It shows that there are hardly any differences between different time periods. The proportion of prepayments increase over time, except during the financial crisis. The time-varying Markov model and GAS model indicate time-variation but fail to show this in terms of outcome since the values are not significant on a 90% confidence level.

Main findings of the research are that higher mortgage interest shrinks partial prepayment in the short term, but increases the likelihood of premature mortgage termination in the long term. The best estimators according to the MNL model are the loan age and the Fico score. Sub-segmentation in partial prepayment does not lead to additional information. Adding a prepayment penalty clause leads to 55% less premature termination, but makes the mortgage less attractive and cannot always be levied. Observing the time-variant models, we conclude that we are not able to improve upon the currently used MNL model in terms of mortgage prepayment estimation. The prepayments show little signs of time-dependency and low interest rate environments do not differ significantly from high interest rate environments. The five-state models suffer from rank deficiency due to the large number of parameters. The two-state models perform better in terms of log-likelihood but lose in terms of errors to the static competitor.

1.2. Problem Description

In the optimal scenario for a mortgagee, the mortgagor pays off the debt according to schedule, which could either be in the form of a Fixed Rate Mortgage (FRM) or Adjustable Rate Mortgage (ARM). In both cases, the mortgagee is able to calculate the rate he needs to charge in order to make the mortgage profitable. However, if a mortgagor decides to refinance or prepay the loan prematurely, the mortgagee is faced with an open position. The mortgagee still needs to pay interest to the financial institution he lend the money from on the long term, but does not earn the expected interest from the home-owner he lend the money to. This financial risk is often referred to as *prepayment risk*. Considering that a mortgagee also needs to take the default risk of mortgagors into account, which is negatively correlated to prepayment risk, he cannot simply minimize this prepayment risk. Because the mortgagee holds the house as a collateral, in most cases prepayment risk is more dangerous than default risk.

The most common form of mortgage prepayment is refinancing. This is generally done whenever the mortgagor finds a better interest rate in the market or when the mortgagor decides to move. One way that is often used in the European market to cover this type of prepayment risk, is to add a prepayment penalty clause to the contract. This penalty clause states that if the borrower makes payments that differ significantly from the payments initially agreed upon in the contract, this will lead to a fine based on a percentage of the remaining mortgage balance or some additional interest for the skipped months.

Soft prepayment penalties allow the mortgagors to sell their house before the contract ends without a penalty. If borrowers decide to refinance their mortgage prematurely, they are faced with a penalty. In case of hard prepayment penalties, the mortgagor is also faced with a penalty if he sells the house before the end of the contract. Some banks levy a prepayment penalty for payments of more than 10% of the original principal per year on top of the contractual payment. However, in the US it is less common to use penalties.

Since prepayment penalties can only lead to additional cost, they are less desirable for mortgagors. Hence, mortgagees can provide more loans by offering contracts without potential fines. However, in that case they still need to deal with the risk of prepayment. On top of that, prepayment penalties only cover a small part of the loss on interest rate as can be seen in Appendix A. Apart from these indications of ineffectiveness it is also unethical or even forbidden by law in some cases. Because a mortgage is usually closed for a period of 20-30 years, a lot of unexpected events can occur in the meantime. For example, one can not give a prepayment penalty to those next of kin when the mortgagor dies. The mortgagee is only able to levy a penalty in case of voluntary prepayment or refinancing.

1.3. Research Question

A convenient way to account for the prepayment risk is to estimate the likelihood of a mortgagor to make a prepayment. This way mortgagees are able to compute the expected loss by means of the probability of prepayment and can determine a client-specific mortgage interest rate. To a large extent, the likelihood of a mortgagor to make a prepayment can be based on a combination of personal characteristics and macroeconomic variables.

Currently banks already build their own models to predict this probability of prepayment, using MNL models, option theoretic models, Markov models, survival analysis and Bayesian methods based on client data from the mortgages they sold. However, since the economic climate is changing, the data may not always be representative for the current situation. The last four years show a positive trend in the housing prices, whereas the savings rate decreases and might even become negative. This is different compared to the situation between 2008 and 2013, where we observe a negative trend for the housing prices as well as the savings rate. In other words, we observe an opposite correlation between these variables in the two different intervals.

Economically, this phenomenon can be explained by the financial crisis. However, clarifying this event does not solve our problem since the economic climate is still different from the situation before the crisis and therefore the old data might not be sufficient. Hence, we need to for new correlations in the data that can explain for both the situation before and after the crisis. In addition to the commonly used explaining variables such as income, loan age, credit score and age, it might be interesting to use latent variables, macroeconomic variables and take indirect effects into account.

For example, the amount of prepayment is correlated with the amount of income, whereas the sign (and size) of this correlation depends on variables such as the mortgage interest rate, the savings rate and possibly other latent variables. Since the savings rate is at an all time low and not yet recovered from the crisis, whereas the housing prices are already above pre-crisis level, this might lead to different relations in comparison with the period before the financial crisis. For this reason, modelling the prepayment risk in a time-varying way could lead to new insights. Adopting latent variables in a Bayesian network might also increase the prediction power of the currently used models. In this research however, we focus mainly on the time-varying aspect, which leads to the following research question:

"*Can we improve upon the mortgage prepayment estimates in low interest rate environments by forecasting mortgage prepayment using a time-varying Markov switching model?*"

2. General Background

This section explains the basic structure of mortgages, the American housing market, macroeconomic development and the way mortgagees deal with the problem of prepayment nowadays in order to get some overall understanding on the topic.

2.1. American Housing Market

According to the Federal Resevere, Central Bank of the US (FED), the outstanding total debt of all mortgage holders on the second quarter of 2017 is equal to 14.6 trillion dollars (Reserve, 2017). Moreover, it is increasing every year over the period 2013-2017. A typical feature of the US housing market is the fact that mortgagors hold the right of voluntary repossession. In this case the borrower can simply hand in the keys of the underlying lien and is no longer obliged to make payments as mentioned in the contract. However, according to Federal Research Economic Data (FRED) only about 2.5% of the single-family residential mortgages are delinquent during a normal economy, which implies that the default rate is even lower (FRED, 2017). As a result of the financial crisis in 2008 this percentage reached a peak of 11.53% in 2010, whereafter it decreased to 2.5% again.

Within a mortgage there are two main components for which the mortgagor pays. The first one is the principal, the initial amount of money that was lent, and the second one the yearly interest rate over this loan. Until the maturity date, which is often 30 years after closing the mortgage, the mortgagor is obliged to make coupon payments every month, consisting of an interest rate part and principal part. If the mortgagor fails to pay, the mortgagee holds a lien on the underlying which they can foreclose. On the other hand, US mortgagors possess the right to file a personal default and hand in the keys of their house at the mortgagee at any time, for example, if the face value of the house is far below the principal value. This is often referred to as an underwater mortgage.

Mortgagors can apply for a mortgage at the bank, an insurer or another big financial institution. Sometimes it is even a possible to get a private loan. The bank or insurer in their turn borrows money from another institution or individual, for example the Federal Reserves, private equity investors or premium deposits. Normally mortgages get packed together in a MBS, which is generally a secured claim on the principal and interest payments based on a set of mortgage loans. A MBS is issued by either a government-sponsored enterprise, a federal government agency or a private financial company. For agency MBSs the payments of principal and interest rate are backed by government guarantee. There are three main agencies that issue and guarantee most of the mortgages; the Federal Home Loan Mortgage Coporation (Freddie Mac), the Federal National Mortgage Association (Fannie Mae) and the Government National Mortgage Association (Ginnie Mae). Together they hold a mortgage debt of 6.95 trillion USD, of which Ginnie Mae holds a debt of 1.85 trillion USD, Freddie Mac of 1.98 trillion USD and Fannie Mae of 3.12 trillion USD. Private mortgage conduits hold an approximated debt of 843 million USD as of 2017Q2. In total 2.87 trillion USD is invested in mortgage pools and trusts, whereas 5.22 trillion USD is invested in federal and related agencies. This means that all in all 8.06 trillion USD is invested in MBS's, implying that 1.14 trillion USD of MBS is owned by others than Fannie Mae, Ginnie Mae and Freddie Mac.

Basically there are two types of mortgages, namely a FRM and an ARM. FRMs have a fixed interest rate over the entire loan term. In case of an ARM the interest rate starts below the market rate and rises over time to eventually overtake the rate of the FRM. In case of high market rates, it gets harder for mortgagors to qualify for a FRM since payments are less affordable. This relation is confirmed by Fisher and Kan (2015) who compare the 30 year FRM rate with the share of ARMs and conclude there is a positive correlation. Since the interest rate is currently around zero, most mortgagors nowadays should rationally hold FRMs, as is indeed the case when we look at Zillow, the leading real estate and rental market place of the US (Zillow, 2018).

Among FRMs we distinguish two types of mortgages which both pay off the entire loan within the total loan term: linear and annuity. In case of a linear mortgage one pays a fixed amount of the principal every month plus the expenses stemming from the interest. This way the mortgagor pays more in the first months in comparison to the last months. In case of an annuity mortgage the mortgagor pays a fixed total amount every month, which consists less and less of interest, since the debt is getting smaller over time. According to Netwerk (2017) only 5.31% of the mortgages are linear, whereas 62.49% are annuity and 30.08% of the market contains of interest-payments-only mortgages. Since interest-payments-only mortgages are not allowed anymore, this type of mortgage will disappear over time.

According to the paper of Badarinza et al. (2017), the percentage of ARM's between 1992-2013 has been between 0.01 and 38.16 percent with a mean of 8.46%. Nowadays only 5.4% of the borrowers chooses an ARM (MotleyFool, 2017). Moench et al. (2010) show that the percentage of ARM's is decreasing over time and present evidence that the ARM share can largely be accounted for by mortgage choice in earlier periods. Hence, we can conclude that the share of FRM's is in general bigger and also depending on the earlier periods.

Furthermore we observe that there exists no culture of prepayment penalties. The new federal mortgage servicing rules state that prepayment penalties can only apply for the first three years and that the amount of penalty is capped. For the first two years, the penalty may not be greater than 2% of the outstanding loan balance. For the third year it is capped at 1% of the outstanding loan (Consumer Financial Protection, 2013). The same report of the Consumer Financial Protection Bureau, an agency of the US government, also states that the mortgagor is permitted to prepay up to 15% of the original principal amount without facing any penalties. This legislation accounts for qualified mortgages that satisfy certain conditions such as a loan term of no more than 30 years and no risky features like an interest-only payments or a negative amortization. For mortgages before 2013 there is no such legislation.

The Federal Housing Finance Agency (FHFA) published a report about the American housing market including forecasts of how the American housing market and the most important macroeconomic variables will develop between 2018 and 2020. According to Lam et al. (2017) the mortgage interest rates, in particular the 30-year fixed rate, will rise again after a historic low of 3.6 percent in 2016 to 5.8 percent in 2020. In addition, it is expected that the unemployment rate, a good estimator for the economy, will remain around 4.6 percent over the next few years. Furthermore the housing market will continue to recover with growths of 4.4, 2.3 and 1.9 percent in 2018, 2019 and 2020, respectively. As a result of the movement in this macroeconomic variables they expect the Housing Affordability Index (HAI), a Moody's Index, to decrease from 167.2 in 2017 to 160.3 in 2020, implying it will be harder for low-income families to buy a home. Despite the fact that the share of mortgage refinancers increased from 39.9 percent in 2014 to 47.4 percent in 2016, due to the low savings rate compared to the rising housing prices, the refinance rate is expected to fall below 20 percent by 2019.

2.2. Main Reasons for Prepayment

Refinancing is, together with relocation, one of the most common forms of prepayment[1] and can to a certain extent be covered by prepayment penalties. This is not always the case though, since the mortgagor is allowed to refinance if he moves to a new house and therefore needs a new mortgage. Also, whenever a mortgagor dies, the next of kin are allowed to sell the house without facing penalties.

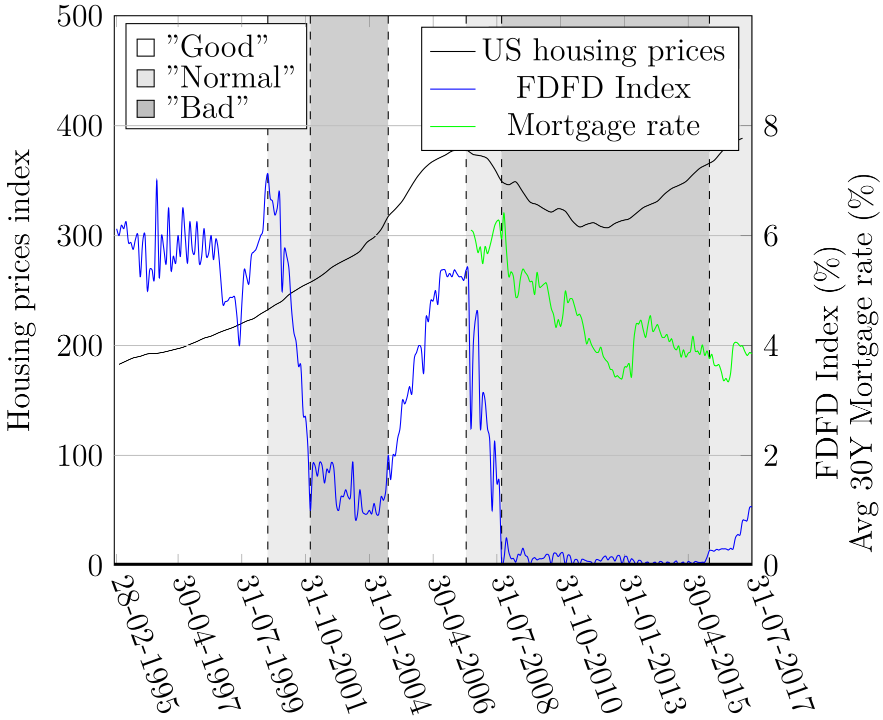
Refinancing of a mortgage is possible, under certain conditions, if the underlying security has a positive overvalue. This is the case whenever the Loan-to-Value (LTV) is lower than one. Since on average the housing prices will rise, this will not cause many troubles. During an event such as the financial crisis however, housing prices tend to decrease, which may lead to underwater mortgages and therefore to additional risks for mortgagors and mortgagees.

2.3. Interest Rates over Time

Obtaining interest rate on savings has always been a steady assumption for most clients that store their savings at a bank. However, the need to save depends on the savings rate, which has never in history been as low as the previous year. The current value around zero is considered an all-time low, as can also be seen in Figure 2.1. Therefore, we speak of a unique situation and we suspect a structural break in the data.

[1] Stated in confidential reports of big banks

Figure 2.1: US Housing prices, FDFD Index & Mortgage rate



In Figure 2.1 data from Bloomberg is used to plot the US housing prices Index versus the Federal Funds Rate (FDFD) Index and the average mortgage rate from end February 1995 until halfway 2017. The shade of grey indicates the state of the economy. As we can observe clearly, the US housing prices drop a little during and after the financial crisis of 2008 from a peak of 378.22 in February 2007 and a trough of 307.12 in June 2012. After June 2012 it starts rising again. The FDFD Index however drops from 5.375% in June 2007 to 0.0625% in September 2008 and starts recovering from January 2017. Besides that, the relative decrease for the FDFD index is much higher than the decrease in housing prices. This tells us that, depending on the mortgage rate, it might be more profitable to make a prepayment after the financial crisis than before 2008.

3. Literature Review

Interest rate risk is one of the many risks that a mortgagee faces (Jaffee, 2003), for which prepayment risk is a specific form. One of the other risks is that the mortgagor defaults, for which the bank still holds a collateral, but is simultaneously faced with an open position. In many articles, such as Bhattacharya et al. (2017), a correlation between default and prepayment is found. In practice mortgage prepayment is mostly estimated with a MNL model and sometimes by survival analysis[1]. In literature however, most research on this topic has been done using survival analysis in combination with a proportional hazard model or with the help of option theoretic models. This section discusses previous studies, including their findings and potential downfalls. In Section 5 we take a more detailed look at the characteristics of the currently used models and compare these with our own Markov models.

Assuming there are low or no prepayment penalties, Kalotay et al. (2004) use an option theoretic model where the mortgagor holds the option to prepay at any time. They show that using this model MBSs and other mortgage pools can be valued. Kelly and Slawson (2001) set up a competing option pricing model with four different prepayment penalties. They find that the value of delaying prepayment is often higher for mortgages with declining-rate penalties than for mortgages with static-rate penalties, since they require a higher interest rate spread to trigger refinancing.

According to Follain et al. (1992) a simple option theoretic model is not adequate to explain aggregate prepayment behaviour compared to a hazard model. Furthermore they investigate mortgage refinancing incentives and find that a decline of 200 basis points on mortgage interest rate often leads to refinancing. Mattey and Wallace (2001) state that mortgage models have difficulties explaining differences in mortgage prepayment among pools with similar interest rates on underlying mortgages. The Freddie Mac database is used to construct an option pricing model plus a hazard model to show that differences in housing price dynamics are an important source of between-pool heterogeneity.

Using survival analysis, it is hard to take time-dependency into account. Dekker et al. (2008) investigate the time-dependent effects and argue that survival analysis fails to take these different effects into account. For example, a mortgagor is less likely to prepay the mortgage in the first few months compared to the last months, implying that the survival rate is higher in the beginning than the end. On top of that, Dekker et al. (2008) state that risk factors may also vary over time, which in our case could be the underlying housing price index or the refinancing rate.

Meis (2015) uses a MNL Model to forecast mortgage prepayment and concludes that the model performs well in terms of accuracy and efficiency. A drawback is that the MNL model assumes all draws to be independent observations. According to Pravinvongvuth and Chen (2005) the biggest downfalls of the MNL model are the fact that it is unable to capture correlation over all paths and the inability to account for perception variance of different paths. The explanation stems from the fact that the random error terms are IID with the same fixed variances (Sheffi, 1985) and the assumption that the covariance matrix of the MNL model is homoskedastic and diagonal (Ben-Akiva et al., 1984). Besides that, the model is unable to capture the effect of variables over time and the presence of indirect or latent effects.

An alternative for the common survival analysis is a hazard model based on a Poisson regression (Schwartz and Torous, 1993). The advantage of this regression is that the grouped data can be used to estimate multiple time scales and non-proportional hazard models, plus it requires less computations. Another possibility is a two-state model, where distinction is made between a segmented and prediction model (Liang and Lin, 2014). Random forest techniques are used to segment mortgagors in different groups, after which a proportional hazard model is constructed to predict the time of prepayment. Liang and Lin (2014) claim that this two-stage model predicts more accurately than a single-stage model without segmentation. There are also examples of Bayesian approaches, like done by Deng and Liu (2009) and I. Popova et al. (2008) but they are less common.

Other findings on mortgage prepayment are that very low-income households are more likely to default and have a lower prepayment probability (Quercia et al., 2012). Ambrose and Sanders (2003) state that the yield curve has a direct impact on the probability of mortgage termination. Furthermore they find no evidence for a relationship between LTV and default or prepayment, which is in contradiction with the results from Meis (2015).

4. Data

4.1. Data Description

We use yearly samples from the Single Family Loan-level Freddie Mac data base, which is publicly available[2]. Each of the samples consists of 50,000 mortgage observations starting in the year the sample is taken from and gets evaluated monthly until maturity or early termination. Between the years 1999-2016, the data contains values like loan-level origination, monthly loan performance and actual loss data on a portion of the fully amortizing 30-year fixed rate Single Family mortgages. Between 2005-2016 it consists of similar data, but now also with a fixed rate for 15-20 year. The data contains mortgage specific values such as the principal, interest rate, loan term, fico score, as well as monthly observations such as the current principal and interest rate. All the specific variables including corresponding explanation can be found in Tables D.1 and D.2 in Appendix D.

[2] `https://freddiemac.embs.com/FLoan/Data/download.php`

4.2. Variable Assumptions and Data Permutation

Since the raw data is unstructured, incomplete and contains several errors, some assumptions are made. To start of, only the years 2001-2016 are used. The reason is that the data from 1999 and 2000 contains too many missing values and errors. Besides that, this data is quite old and does not contain special values on low interest rate periods or other occasions that do not occur between 2001 and 2016. Furthermore the first seven observations in the monthly performance file are not taken into account since it is stated in the user guide that the first six months are mostly incorrect. Since the first performance observations consist mostly of loan age equal to zero, we chose to ignore the first seven observations.

Once a mortgage hits the delinquency level of six months, we define it as default. The remainder of observations of that mortgagor are deleted from the database. If the delinquency status is REO dispositioned and does not fit one of the other prepayment states, the observation is deleted from the database. Since the file contains only fixed rate fully amortizing mortgages, the monthly prepayment is calculated as follows:

$$C = P^{orig}\frac{r^{mo}(1+r^{mo})^n}{(1+r^{mo})^n-1}, \tag{4.1}$$

where C represents the coupon, measured by the amount of payment per month(€/month). P^{orig} is the Original Principal (€) and P^{cur} is the Current outstanding Principal (€). r^{mo} stands for interest rate per month (%). We assume that the monthly interest can be compounded using the annual rate r^{an} by means of $r^{mo} = (1+r^{an})^{1/12} - 1$. Finally, n presents the number of periods (months). In case of 30 years this is equal to 360.

The monthly coupon C is the same over time and consists of an interest part (C_t^{int}) and a principal part (C_t^{prin}) of which the ratio $\frac{C_t^{prin}}{C_t^{int}}$ increases as time t increases. The Interest and Principal part can be calculated by

$$C_t^{int} = P_{t-1}^{cur} r^{mo}, \qquad C_t^{prin} = C - C_t^{int}. \tag{4.2, 3}$$

4.3. Summary Statistics

Now that we identified the variables in both the origination file and the monthly performance file, we are able to obtain some summary statistics. This paragraph contains the summary statistics of some important variables of the origination file and monthly performance file between 2001-2016. The complete summary statistics can be found in Tables E.1 and E.2 in Appendix E.

As we can see clearly from Table 4.1a, over 96% of the mortgages is prepayed. Furthermore Table 4.1b shows that 93.4% of the Fico scores are higher than 650, indicating that they have a good credit history. Figure 4.3 tells us that the height observations of Fico Scores are upward sloping, indicating that lower Fico scores are rare in our dataset. Combining these two findings makes this dataset suitable for our research. It tells us that most mortgagors are likely to pay their bills and almost all of them prepay.

Table 4.1: Summary statistics Reason mortgage ending & Fico Score

(a) Reason for end mortgage (Zero Balance Code)

Reason	# Obs.	Percentage
Prepayed	495,288	96.08
Foreclosed	7,186	1.39
Repurchase	2,006	0.39
REO Disposition	11,028	2.14

(b) Fico Score

Fico	# Obs.	Percentage
NA	679	0.09
0-600	7,393	0.95
600-650	42,972	5.54
650-700	122,912	15.86
700-750	203,682	26.28
750-800	317,528	40.97
>800	79,834	10.30

Table 4.2: Summary Statistics Debt-to-Income & Loan-To-Value

(a) Debt to income (%)

DTI	# Obs.	Percentage
NA	9,147	1.18
0-15	33,131	4.27
15-20	53,607	6.92
20-25	85,041	10.97
25-30	106,594	13.75
30-35	118,089	15.24
35-40	121,328	15.66
40-45	117,400	15.15
45-50	75,687	9.77
>50	54,976	7.09

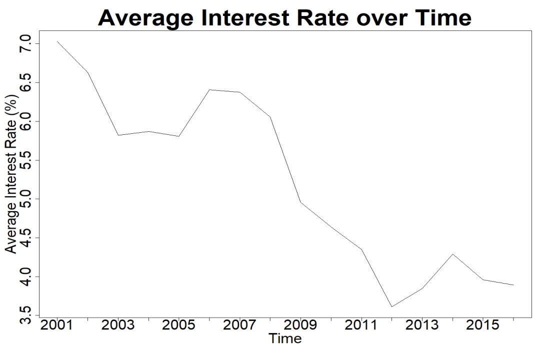
(b) LTV

LTV	# Obs.	Percentage
NA	25	0
0-40	49,016	6.32
40-60	120,735	15.58
60-70	102,632	13.24
70-74	57,969	7.48
74-76	52,756	6.81
76-80	61,927	7.99
80-90	225,007	29.03
>90	104,933	13.54

Table 4.2a contains insight on the number of observations per segment of Debt-to-Income (DTI). Taking a closer look to the distribution as shown in Figure 4.3 we observe that these observations seem normally distributed around a mean of 35% which is slightly skewed to the right. Table 4.2b shows a steep peak for mortgages with a LTV around 80%. Closer inspection in Figure 4.3 reveals that the mode of the observations is indeed

equal to exactly 80%. This is due to the fact that conforming loan guidelines state that the LTV ratio must be less or equal to 80%. In other words, this is the maximum amount mortgagors can get without losing favourable mortgage characteristics.

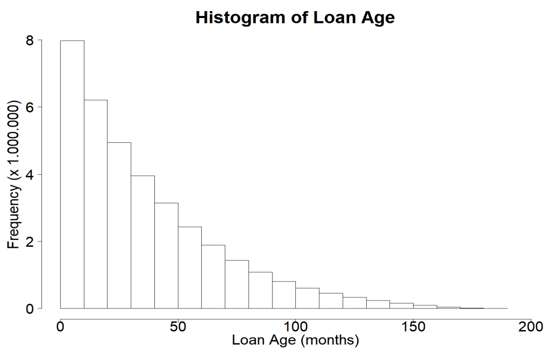
Looking at delinquency, we observe that in more than 96% of the cases there is no delinquency and that in about 2% of the cases the delinquency is more than one month. This is in accordance with the results of Table 4.1a which also show that most of the mortgages are prepayed. Finally from Table 3.4 we observe that 66.36% of the interest rates lie below 6%. Furthermore Figure 4.1 shows the average interest rate decreases over time, which can be explained by the fact that the savings rate also decreases over time as we have seen in Figure 2.1.



Interest Rate

Interest rate (%)	# Obs.	Percentage
0-5	331,762	42.81
5-6	182,540	23.55
6-6.25	54,873	7.08
6.26-6.5	61,945	7.99
6.5-6.75	45,666	5.89
6.75-7	46,543	6.01
7-7.5	41,439	5.35
7.5-8	8,186	1.06
8-8.5	1,465	0.19
>8.5	581	0.07

Figure 4.1 & Table 4.4: Average Interest Rate

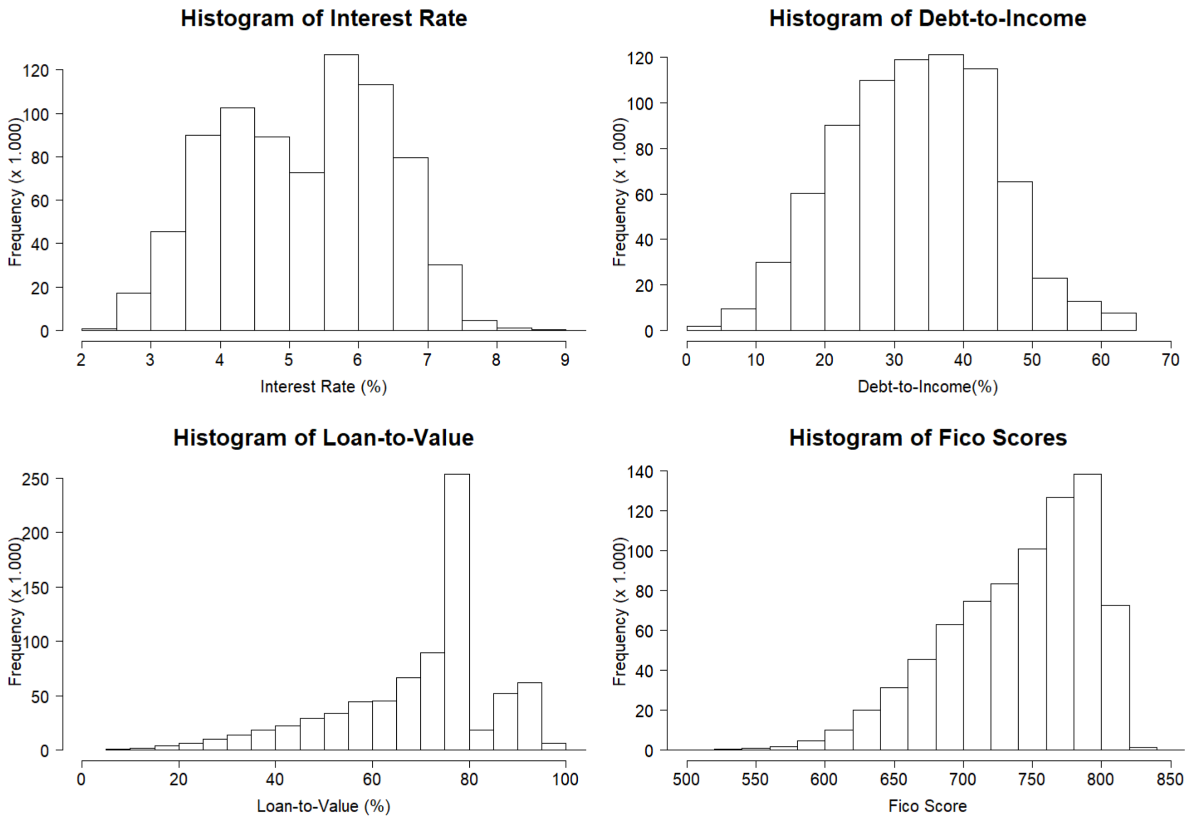


Loan Age

Months	# Obs.	Percentage
0-12	9,323,175	26.03
13-24	6,975,260	19.47
25-36	5,302,963	14.80
37-60	7,060,813	19.71
61-120	6,274,854	17.52
>120	885,861	2.47

Figure 4.2 & Table 4.5: Loan Age

Figure 4.3: Summary Statistics 2001-2016



Overall we can conclude that, based on the summary statistics that we have found, the dataset seems normal and suitable for our research. There are lots of mortgagors that prepay, the dataset captures the lowering savings rate by means of interest rates and the other variables such as loan age, LTV, DTI and Fico are distributed as expected. Therefore we conclude that the dataset fits our needs.

Table 4.3: Delinquency Status (months)

Delinquency	# Obs.	Percentage
0*	34, 571, 695	96.51
1	526, 962	1.47
2-3	235, 805	0.66
4-6	100, 756	0.28
7-12	181, 203	0.51
13-24	108, 262	0.30
>24	75, 985	0.21
REO** Acquired	22, 258	0.06

* Incl. Prepayments, **REO= Real Estate Owned

5. Methodology

In this section we start by making a segmentation in different states of mortgage risk, whereafter we discuss the currently used models, their downfalls and suggest alternatives to correct for these imperfections. As mentioned earlier in Section 3, the frequently used models are the MNL Model and Survival Analysis. As new method we suggest a Time-Varying Markov Switching Model to correct for time-dependency. Since macroeconomic variables differ during different economic regimes, we suspect that the state of the economy, and hence time, has a big effect on mortgage prepayment.

5.1. Subclasses of Risk

First of all we distinguish several types of risk. As stated in Section 1.2, default and delinquency can also form a risk for the mortgagee. Besides that, we are able to separate different levels of prepayment. To evaluate the risk, we divide the risk in the following different states:

$$Y_{it} = \begin{cases} k = 1 : \text{On Schedule/Contract}, & \text{if} \quad \Delta_{it} \approx 0, \\ k = 2 : \text{Default}, & \text{if} \quad \Delta_{it} = \text{Def}_i, \\ k = 3 : \text{Delinquent}, & \text{if} \quad \text{Def}_i < \Delta_{it} < 0, \\ k = 4 : \text{Partial Prepayment}, & \text{if} \quad 0 < \Delta_{it} < \text{Prep}_{it}, \\ k = 5 : \text{Full Prepayment}, & \text{if} \quad \Delta_{it} = \text{Prep}_{it}, \end{cases} \tag{5.1}$$

where Y_{it} indicates the mortgage state of mortgagor i at time t. Δ_{it} indicates the payment deviation from the contract of mortgagor i at time t. Def_i is the maximum number of delayed months for which mortgagor i can be declared as default and fixed to six months, whereas Prep_{it} indicates the total outstanding principal according to contract for mortgagor i on time t.

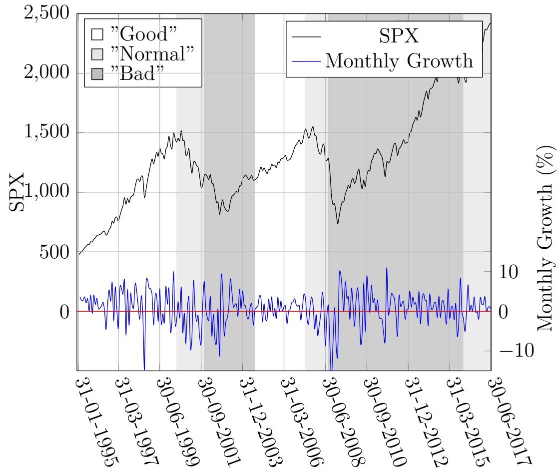
Since we are interested in prepayment risk specifically, we could also split up the 'Partial Prepayment' state even more in such way that we obtain the following possible states where the prepayments are split up in equal segments Prep^{lev} of 10% each:

$$
Y_{it} = \begin{cases} k = 1 : \text{On Schedule/Contract}, & \text{if} \quad \Delta_{it} \approx 0, \\ k = 2 : \text{Default}, & \text{if} \quad \Delta_{it} = \text{Def}_i, \\ k = 3 : \text{Delinquent}, & \text{if} \quad Def_i < \Delta_{it} < 0, \\ k = 4.1 : \text{Partial Prepayment 1}, & \text{if} \quad 0 < \Delta_{it} < \text{Prep}_{it}^{lev1}, \\ k = 4.2 : \text{Partial Prepayment 2}, & \text{if} \quad \text{Prep}_{it}^{lev1} < \Delta_{it} < \text{Prep}_{it}^{lev2}, \\ \qquad \vdots & \quad \vdots \qquad \vdots \\ k = 4.9 : \text{Partial Prepayment 9}, & \text{if} \quad \text{Prep}_{it}^{lev8} < \Delta_{it} < \text{Prep}_{it}^{lev9}, \\ k = 4.10 : \text{Partial Prepayment 10}, & \text{if} \quad \text{Prep}_{it}^{lev9} < \Delta_{it} < Prep_{it}, \\ k = 5 : \text{Full Prepayment}, & \text{if} \quad \Delta_{it} = Prep_{it}. \end{cases} \tag{5.2}
$$

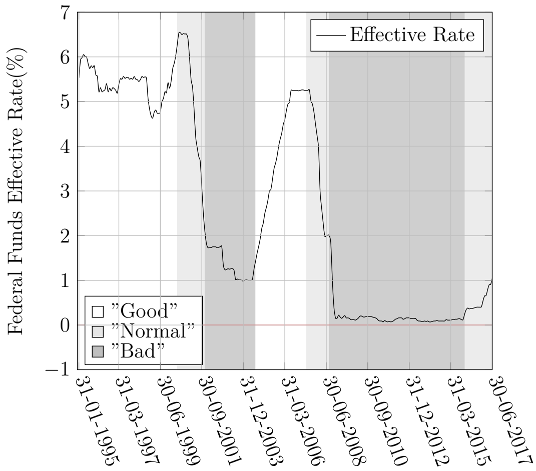
In addition we look for possibilities to model the problem in a time variant way. This is done by modelling the problem such that we can not only switch between prepayment states overall, but we set assume different probabilities in different times in terms of 3 states of the economy $t_{ec} = 1, 2, 3$ for which:

$$
Y_{it} = \begin{cases} t_{ec} = 1 : \text{Recession}, & \text{if} \quad \text{Growth}_{\text{Quarterly}}^{\text{SPX}} < 0\%, \\ t_{ec} = 2 : \text{Normal}, & \text{if} \quad 0 \leq \text{Growth}_{\text{Quarterly}}^{\text{SPX}} < 2\%, \\ t_{ec} = 3 : \text{Expansion}, & \text{if} \quad 2\% \leq \text{Growth}_{\text{Quarterly}}^{\text{SPX}}. \end{cases} \tag{5.3}
$$

Both the SPX and the Effective seem good estimators for the state of the economy. Above in Equation (5.3) the SPX is used as an example. To illustrate their relevance and contradictory process during the bad state of the economy, the monthly Growth of the SPX and the monthly Effective Rate of the FED are shown in Figures 5.1a and 5.1b respectively.



(a) Monthly SPX & Growth 1995-2017



(b) Monthly Effective Rate 1995-2017

5.2. Multinomial Logit Model

Most banks use a MNL model to forecast mortgage prepayment[3], which can be modelled in a way that has also been done by Vasconcelos (2010) and Meis (2015). A MNL Model basically measures the relative probability of being in one state compared to another. In our case this is done in the following way.

For i = mortgagor $1, \ldots, N$ and $k = 1, \ldots, K$ possible payment states, we set up the following linear predictor model, such that over all observations on time t:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i} = \boldsymbol{X}_i'\boldsymbol{\beta}_k.$$

Next, for time $t = 1, \ldots, T$, we set the independent binary regressions equal to

$$\ln\left(\frac{\mathbb{P}[Y_{it} = k]}{\mathbb{P}[Y_{it} = K]}\right) = \boldsymbol{X}_{it}'\boldsymbol{\beta}_k. \tag{5.4}$$

Based on Equation (5.4), we back out $\mathbb{P}[Y_{it} = k]$ as shown in Equation (5.5). Using all the values of k, given by $j = 1, \ldots, K-1$, it is possible to calculate $\mathbb{P}[Y_{it} = K]$ for every i and t, as shown in Equation (5.6).

$$\mathbb{P}[Y_{it} = k] = \mathbb{P}[Y_{it} = K]e^{\boldsymbol{X}_{it}'\boldsymbol{\beta}_k}, \qquad \mathbb{P}[Y_{it} = K] = 1 - \sum_{k=1}^{K-1} \mathbb{P}[Y_{it} = K]e^{\boldsymbol{X}_{it}'\boldsymbol{\beta}_k}. \tag{5.5, 6}$$

Note that by dividing the left and right side of Equation (5.6) by a factor $\mathbb{P}[Y_{it} = K]$, it can be re-written as Equation (5.7) and hence we can calculate $\mathbb{P}[Y_{it} = k]$ by implementing it in Equation (5.5):

[3]Source: Big banks, source not allowed due to confidentiality

$$\mathbb{P}[Y_{it} = K] = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\boldsymbol{X}'_{it}\beta_j}}, \qquad \mathbb{P}[Y_{it} = k] = \frac{e^{\boldsymbol{X}_{it}\beta_k}}{1 + \sum_{j=1}^{K-1} e^{\boldsymbol{X}'_{it}\beta_j}}. \tag{5.7, 8}$$

Eventually we are able to compute a log-likelihood for all mortgagors i over all time periods t and possible states k of the MNL model in the following way:

$$\log L(\boldsymbol{\beta}) = \sum_i \sum_k \sum_t \ln \mathbb{P}[Y_{it} = k] \cdot \mathbb{1}\{Y_{it} = k\}. \tag{5.9}$$

Here $\mathbb{1}\{Y_{it} = k\}$ is an indicator function, depending on the payment state of Y_{it}. The function takes value 1 if k is equal to the payment state of mortgagor i on any time t and 0 otherwise. Based on the log-likelihood the Akaike Information Criterion (AIC) is computed in order to compare each separate MNL model. More information about the AIC can be found in Appendix C.

The big advantage of the MNL Model is that it is easy to implement. Besides the inability to capture correlation over all paths, as explained in Section 3, another big disadvantage is the risk of overfitting. The multinomial logit regression is vulnerable to overconfidence. For example, our dataset might contain certain values for which every case ends up in prepayment, whereas in fact there are of course always cases for which this is not necessarily the case. In other words, there is a good probability of bias. The more observations, the smaller this probability of bias. Since we posses almost 30 million observations, we ignore this risk.

5.3. Survival Analysis

The general survival and lifetime functions are given below in Equation (5.10) and (5.11) as stated by Mills (2010), where $S_k(t)$ indicates the survival function and T the time until death or in our case termination of the mortgage. $F_k(t)$ reflects the lifetime distribution function and is directly related to the survival function.

$$S_k(t) = \mathbb{P}[T_k > t], \qquad F_k(t) = \mathbb{P}[T_k \leq t] = 1 - S_k(t). \tag{5.10, 11}$$

Based on the above two distributions we are now able to compute $s_k(t)$, the so called event density function:

$$s_k(t) = \frac{\delta}{\delta t} S_k(t) = \frac{\delta}{\delta t}(1 - F_k(t)) = -f_k(t).$$

The hazard rate for payment state k, denoted by $\lambda_k(t)$, is the probability that a mortgagor in state k transfers to another state at time t, given that it has not switched states before t. In the papers of Schwartz and Torous (1989) and Meis (2015) this hazard rate is referred to as the prepayment function and is stated as follows:

$$\begin{aligned} \lambda_k(t,x) &= \lim_{\delta t \to 0} \frac{\mathbb{P}[T_k < t + \delta t | T_k \geq t]}{\delta t} \\ &= \lim_{\delta t \to 0} \frac{\mathbb{P}[t \leq T_k < t + \delta t]}{\delta t \cdot S_k(t)} \\ &= \frac{f_k(t)}{S_k(t)} = -\frac{s_k(t)}{S_k(t)} \equiv \lambda_{0k}(t) e^{\boldsymbol{x}'_{it}\boldsymbol{\beta}_k}. \end{aligned} \tag{5.12}$$

The last step is based on the logistic model by Cox (1992) as mentioned by Rodrıguez (2005), which transforms the model to discrete time:

$$\lambda(t,x) = \lambda_0(t) e^{\boldsymbol{x}'\boldsymbol{\beta}}, \qquad \frac{\lambda(t,x)}{1-\lambda(t,x)} = \frac{\lambda_0(t)}{1-\lambda_0(t)} e^{\boldsymbol{x}'\boldsymbol{\beta}}. \tag{5.13, 14}$$

Based on the hazard rate, we identify the cumulative hazard rate $\Lambda_k(t)$ by integrating the form of the hazard rate on the fourth line of Equation (5.12). With this cumulative hazard rate we can back out $\lambda_k(t)$

$$\Lambda_k(t) = -\ln S_k(t) \qquad \rightarrow \qquad S_k(t) = e^{-\Lambda_k(t)},$$

$$\Lambda_k(t) = \int \lambda_k(u) du = \sum_i \sum_t \lambda_k(t).$$

Using the third line of Equation (5.12), we obtain $f_k(t) = \lambda_k(t) S_k(t)$. Eventually, it is possible to once again calculate the log-likelihood similar to Equation (5.9), i.e.

$$\log L = \sum_i \sum_k \sum_t \ln f_k(t) \cdot \mathbb{1}\{Y_{it} = k\}, \tag{5.15}$$

where $\mathbb{1}\{Y_{it} = k\}$ is an indicator function, depending on the payment state of $f_k(t)$. The function takes value 1 if k is equal to the payment state of mortgagor i on any time t and 0 otherwise.

5.4. Regular Markov Switching Model

Another way to model the prepayment risk is by means of a Markov switching model. We do so by modelling the states as described in Equation (5.1). This way we are able to estimate the transition probabilities between these states. Note that state 2 (Default) and state 5 (Full Prepayment) are absorbing states which means that once we enter those states, we will not leave them later on. States 1,3 and 4 are all transient, which means we are not certain to return in the near or far future.

The maximum likelihood estimators of the transition possibilities for every month t and the corresponding standard errors are calculated as shown below in Equations (5.16) and (5.17).

$$\hat{\pi}_{ij,t}^{MLE} = \frac{n_{ij,t}}{\sum_{k=1}^{K} n_{ik,t}}, \qquad \hat{\sigma}(\hat{\pi}_{ij,t}^{MLE}) = \frac{\hat{\pi}_{ij,t}^{MLE}}{\sqrt{n_{ij,t}}}, \tag{5.16, 17}$$

where n_{ij} equals the number of observations going from state i to j for $i, j = 1, \ldots, 5$. Equations (5.18) and (5.19) show the transition probability matrix and the corresponding standard errors. Here $\hat{\pi}_{22} = \hat{\pi}_{55} = 1$ and $\hat{\pi}_{2j} = \hat{\pi}_{5j} = 0$ for all other values of j.

$$\hat{\mathbf{\Pi}} = \begin{array}{c} \\ Y_{1,t+1} \\ \vdots \\ Y_{5,t+1} \end{array} \begin{array}{c} \begin{array}{ccc} Y_{1,t} & \dots & Y_{5,t} \end{array} \\ \begin{bmatrix} \pi_{11} & \dots & \pi_{51} \\ \vdots & \ddots & \vdots \\ \pi_{15} & \dots & \pi_{55} \end{bmatrix} \end{array}, \qquad \hat{\mathbf{\Sigma}} = \begin{array}{c} \\ Y_{1,t+1} \\ \vdots \\ Y_{5,t+1} \end{array} \begin{array}{c} \begin{array}{ccc} Y_{1,t} & \dots & Y_{5,t} \end{array} \\ \begin{bmatrix} \sigma_{11} & \dots & \sigma_{51} \\ \vdots & \ddots & \vdots \\ \sigma_{15} & \dots & \sigma_{55} \end{bmatrix} \end{array}. \tag{5.18, 19}$$

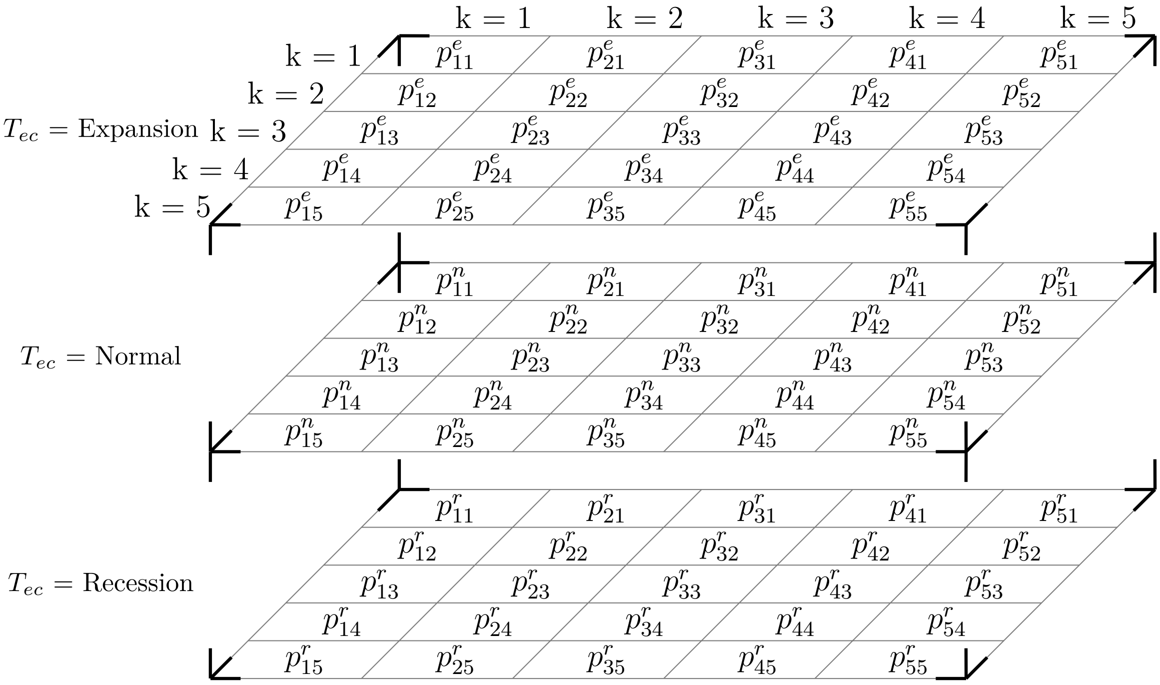
The log-likelihood of the Markov chain can be calculated as follows:

$$\log L(\mathbf{\Pi}) = \sum_{i} \sum_{j} n_{ij} \log(\pi_{ij}). \tag{5.20}$$

5.5. Time-Varying Markov Switching Model

In a regular Markov Switching Model we assume that the transition probabilities are the same at all time. Since in practice we want to take the changing economic climate into account, we want to construct a model where we basically set up multiple Markov Switching Models, for which there exist transitions probabilities between different payment states and different economic states. Below in Figure 5.2 the visualisation of the time varying Markov switching model is shown.

Figure 5.2: Visualisation of Transition matrix in over 3 Economic states



To model this representation in practice, we make use of the paper written by Bazzi et al. (2017), which describes methods to model a time-varying Markov Switching Model. The mathematical way they model it in their paper is stated below in Section 5.5.1. A big advantage of this approach is that it is able to give a more precise estimate for different time intervals, which results in less bias. The trade-off here is that we need to estimate more parameters and therefore a less accuracy in terms of variance.

5.5.1. Technical Representation

In case of the regular Markov model we calculate the chance of going from one payment state to another. This chance of going from state i to state j, given by π_{ij}, is equal to the $(i+1,\, j+1)^{th}$ element of a $K \times K$ matrix $\mathbf{\Pi}$. For all the non-negative elements π_{ij}, where z_t is a hidden discrete process, it holds that:

$$\pi_{ij} = \mathbb{P}[z_t = j | z_{t-1} = i], \qquad \sum_{j=1}^{K} \pi_{ij} = 1, \qquad \pi_{ij} \leq 0, \qquad \forall i, j \in \{1, \ldots, K\}. \qquad (5.21)$$

The conditional density of y_t given $\boldsymbol{\psi}$, where $\boldsymbol{\psi} = (\boldsymbol{\sigma}^2, \mathbf{\Pi})'$, and all other information on time $t-1$, stated by I_{t-1}, for joint stochastic process $\{z_t, y_t\}$ is given by

$$p(y_t|\boldsymbol{\psi}, I_{t-1}) = \sum_{i=1}^{K} p(y_t|\theta_i, \boldsymbol{\psi})\mathbb{P}[z_t = i|\boldsymbol{\psi}, I_{t-1}] = \sum_{i=1}^{K}\sum_{k=1}^{K} p(y_t|\theta_i, \boldsymbol{\psi}) \cdot \pi_{ki} \cdot \mathbb{P}[z_{t-1} = k|\boldsymbol{\psi}, I_{t-1}], \qquad (5.22)$$

where all parameters $\boldsymbol{\psi}$ and $\theta_1, \ldots, \theta_K = \mu_1, \ldots, \mu_K$ are unknown.

It is possible to rewrite this expression in matrix notation by defining $\boldsymbol{\xi}'_{t-1}$ as the a $K \times 1$ vector containing all probabilities $\mathbb{P}[z_t = i|\boldsymbol{\psi}, I_{t-1}]$ and $\boldsymbol{\eta}_t$ a K-dimensional vector of densities $p(y_t|\theta_i, \boldsymbol{\psi})$ for $i = 1, \ldots, K$. Hence, Equation (5.22) simplifies to:

$$p(y_t|\boldsymbol{\psi}, I_{t-1}) = \boldsymbol{\xi}'_{t-1}\boldsymbol{\Pi}\boldsymbol{\eta}_t. \tag{5.23}$$

Using the Hamilton recursion and the Hadamard element ($\odot$) we can update $\boldsymbol{\xi}_t$ such that

$$\boldsymbol{\xi}_t = \frac{(\boldsymbol{\Pi}'\boldsymbol{\xi}_{t-1}) \odot \boldsymbol{\eta}_t}{\boldsymbol{\xi}'_{t-1}\boldsymbol{\Pi}\boldsymbol{\eta}_t}. \tag{5.24}$$

In order to build a Markov Model with time varying transition probabilities (Bazzi et al., 2017) a dynamic parameter vector $\boldsymbol{f}_t$ is introduced by separating it from the parameter $\boldsymbol{\psi}$, which leaves us with a static parameter $\boldsymbol{\psi}^* = (\sigma^2, \boldsymbol{\omega}, \boldsymbol{A}, \boldsymbol{B})$. This way we can update the obtained dynamic parameter in the following way:

$$\boldsymbol{f}_{t+1} = \boldsymbol{\omega} + \boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{f}_t, \qquad \boldsymbol{s}_t = \boldsymbol{S}_t \cdot \boldsymbol{\nabla}_t, \qquad \boldsymbol{\nabla}_t = \frac{\delta}{\delta \boldsymbol{f}_t} \log p(y_t|\boldsymbol{f}_t, \boldsymbol{\psi}^*, I_{t-1}). \tag{5.25}$$

Here $\boldsymbol{\omega}$ is a constant, $\boldsymbol{A}$ and $\boldsymbol{B}$ are coefficient matrices and $\boldsymbol{s}_t$ is the scaled score of the predictive observation density with respect to $\boldsymbol{f}_t$ using the scaling matrix $\boldsymbol{S}_t$.

For two states it holds that

$$\boldsymbol{\nabla}_t = \frac{p(y_t|\theta_0, \boldsymbol{\psi}^*) - p(y_t|\theta_1, \boldsymbol{\psi}^*)}{p(y_t|\boldsymbol{\psi}^*, I_{t-1})} g(\boldsymbol{f}_t, \boldsymbol{\psi}^*, I_{t-1}), \tag{5.26}$$

$$g(\boldsymbol{f}_t, \boldsymbol{\psi}^*, I_{t-1}) = \begin{pmatrix} \mathbb{P}[z_{t-1} = 0|\boldsymbol{\psi}^*, I_{t-1}] \cdot (1 - 2\delta_{00})\pi_{00,t}(1 - \pi_{00,t}) \\ -\mathbb{P}[z_{t-1} = 0|\boldsymbol{\psi}^*, I_{t-1}] \cdot (1 - 2\delta_{11})\pi_{11,t}(1 - \pi_{11,t}) \end{pmatrix}. \tag{5.27}$$

In a similar way, with $i = 1, \ldots, K$ and $j = 1, \ldots, K-1$, we can obtain time varying transition probabilities for K states

$$\pi_{ij,t} = \delta_{ij} + (1 - 2\delta_{ij}) \exp(f_{ij,t}) \left(1 + \sum_{j=1}^{K-1} \exp(f_{ij,t})\right)^{-1}, \quad \pi_{i,K-1,t} = 1 - \sum_{j=1}^{K-1} \pi_{ij,t}(\delta_{ij}). \tag{5.28, 29}$$

Here, $f_{ij,t}$ are the time varying parameters corresponding to the time varying transition probabilities $\pi_{ij,t}$ and collected in a $K(K-1) \times 1$ vector $\boldsymbol{f}_t$ which can be updated like in Equation (5.25)

$$\boldsymbol{\nabla}_t = \boldsymbol{J}'_t \boldsymbol{\nabla}^{\Pi}_t, \qquad \mathcal{I}_{t-1} = \mathbb{E}[\boldsymbol{J}'_t \boldsymbol{\nabla}^{\Pi}_t \boldsymbol{\nabla}^{\Pi\prime}_t \boldsymbol{J}_t], \tag{5.30, 31}$$

$$\boldsymbol{\nabla}^{\Pi}_t = \frac{\partial \log p(y_t|\boldsymbol{\psi}^*, I_{t-1})}{\partial \mathrm{vec}(\boldsymbol{\Pi})'} = \frac{\boldsymbol{\eta}_t \otimes \boldsymbol{\xi}_{t-1}}{p(y_t|\boldsymbol{\psi}^*, I_{t-1})}. \tag{5.32}$$

The element $\boldsymbol{J}_t = \frac{\partial \text{vec}(\boldsymbol{\Pi}_t)}{\partial \boldsymbol{f}_t'} = \frac{\delta \pi_{ij,t}}{\delta f_{i'j',t}}$ is given by

$$\frac{\delta \pi_{ij,t}}{\delta f_{i'j',t}} = \begin{cases} (1-2\delta_{ij})\pi_{ij,t}(1-\pi_{ij,t}), & \text{for } i = i' \wedge j = j', \\ -(1-2\delta_{ij})\pi_{ij,t}\pi_{ij',t}, & \text{for } i = i' \wedge j \neq j', \\ 0, & \text{otherwise.} \end{cases} \tag{5.33}$$

The log-likelihood is eventually calculated by:

$$\log L(\boldsymbol{\psi}) = \sum_t \log(\boldsymbol{\xi}_{t-1}' \boldsymbol{\eta}_t). \tag{5.34}$$

Estimating both the $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices in (5.25) leads to the full generalized autoregressive score (GAS) model as described by Bazzi et al. (2017), whereas fixing the $\boldsymbol{B}$ to an identity matrix leads Time-Varying Probabilities (TVP) model. Fixing both $\boldsymbol{B}$ to and identity matrix and $\boldsymbol{A}$ to a zero matrix leads to a static Markov Switching model.

6. Results

6.1. MNL Model Setup and Preliminary Probabilities

This paragraph contains the first results of the MNL model as ran in the following way for every single mortgagor i:

$$\ln\left(\frac{\mathbb{P}[k=\kappa]}{\mathbb{P}[k=1]}\right) = \alpha_\kappa + x_{1i}\beta_{1\kappa} + x_{2i}\beta_{2\kappa} + \cdots + x_{Mi}\beta_{M\kappa}, \text{ for } \kappa = 2, \dots, 5, \tag{6.1}$$

where $x_{1i}, \dots, x_{Mi}$ are the explaining variables for mortgagor i as shown in Table D.1 in Appendix D and $\beta_{1\kappa}, \dots, \beta_{M\kappa}$ the corresponding regression coefficients. Such that:

$$\ln\left(\frac{\mathbb{P}[\text{Default}]}{\mathbb{P}[\text{On Schedule}]}\right) = \alpha_2 + \text{LoanAge}_i\beta_{12} + \text{DelSts}_i\beta_{22} + \cdots,$$
$$\vdots$$
$$\ln\left(\frac{\mathbb{P}[\text{Full Prepayment}]}{\mathbb{P}[\text{On Schedule}]}\right) = \alpha_5 + \text{LoanAge}_i\beta_{15} + \text{DelSts}_i\beta_{25} + \cdots.$$

We choose only to take care of variables of the origination file at first, since these variables are known at the closure of a mortgage. As a benchmark we took the MNL Model with the divisions as stated in Equation (5.1). The results of this model are shown in Table 6.1a. Next we extended this model to take a closer look at the prepayment cases, such as shown in Equation (5.2). The results are shown in Table 6.1b. We observe that of the mortgages that are partly prepayed, 83.9% is in state 4.1, meaning that the prepayment is less than 10% of the outstanding principal.

Table 6.1: Percentage of times being in Payment State k

(a) No subsegmentation

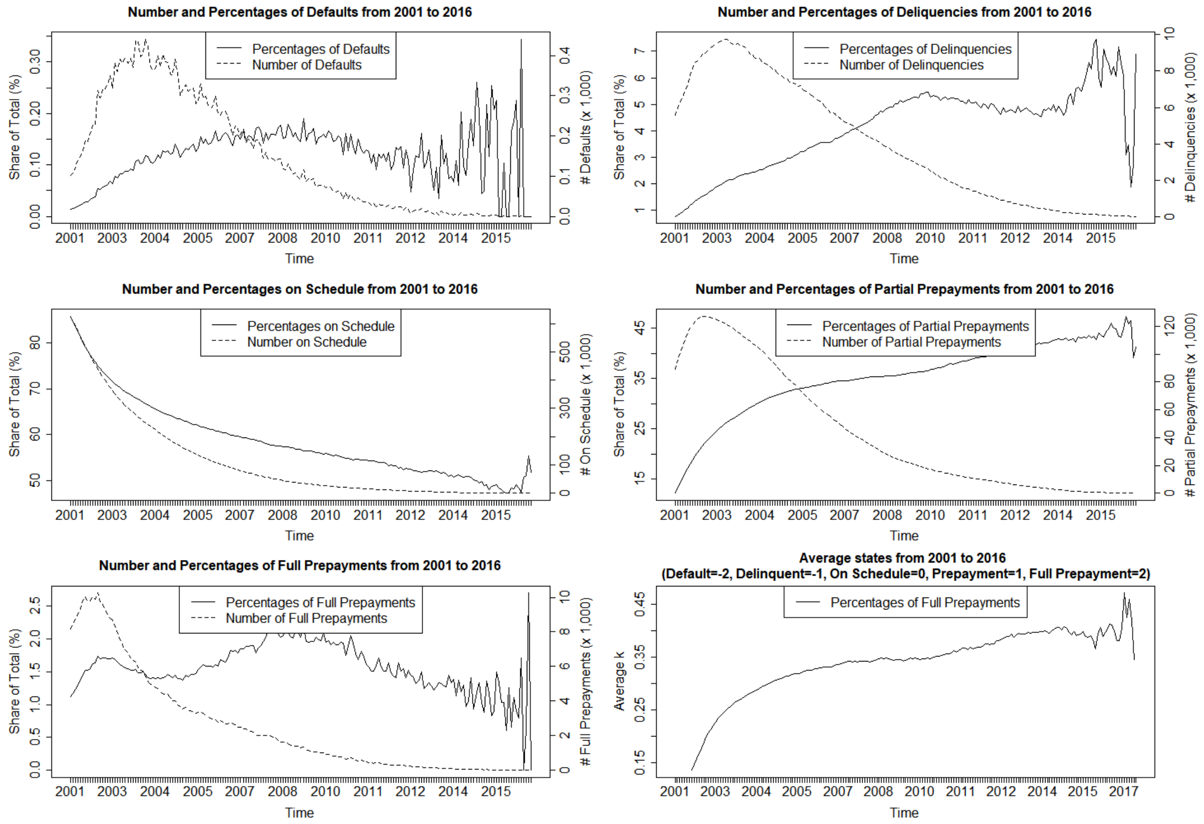
k	# Obs.	Percentage
1	20,304,252	69.04
2	26,607	0.09
3	722,130	2.46
4	7,893,072	26.84
5	464,427	1.58
Total:	29,410,448	100

(b) subsegmentation in k=4

k	# Obs.	Percentage
4.1	6,625,870	22.53
4.2	592,840	2.02
4.3	226,907	0.77
4.4	134,253	0.46
4.5	94,888	0.32
4.6	72,075	0.25
4.7	54,253	0.18
4.8	42,092	0.14
4.9	30,849	0.10
4.10	19,045	0.06
Total:	7,893,072	26.84

Furthermore, Figure 6.1 shows a plot of the number and percentages of mortgagors in every state for every month to check whether their exist peaks in certain states for specific times. Because the absolute number of observations in a certain state does not give us all the information, the share of observations in state k is also plotted. We can see clearly see a decrease in the observations on schedule over time. This can be explained by the fact that every mortgage starts on schedule, but can differ more over time. In first instance it seems that the number of delinquencies and prepayments also decrease over time, but the percentage shows us that in fact they both increase and stagnate.

Figure 6.1: Numbers and Percentages of being in state k for $k = 1, \ldots, 5$ from 2001-2016



6.1.1. Results MNL Models

Several explaining models have been tried of which the most important ones can be found in Appendix F. Table F.1 shows the results of the MNL model for all variables in the origination file of 2001 except for the superconforming flag since this value was zero and not significant on any level. This is done to give us an indication of which variables we should use for our model. Table F.2 shows the results with all explaining variables minus the superconforming flag and the original Unpaid Principal Balance as this variable had very little explaining power. We observed that the Prepayment penalty has the largest effect per unit increase, but need to bear in mind that this parameter is a binary variable and therefore can only move with one unit, which explains that the effect per unit is bigger.

Next we observe that the respectively the Interest Rate, First Home indicator, Number of Borrowers, Number of Units and Loan term have the biggest effects. The Interest rate is an important one, since this can fluctuate some percentages and is easily adjustable over time. The rest of these variables are less important since they are mainly dummies. Among the parameters with a wide range, the FICO score and DTI explain good, even though DTI is not significant for state 2. Furthermore the LTV has more explaining power than the CLTV. Hence we choose to incorporate the Prepayment Penalty, FICO, DTI and LTV as time constant parameters.

For the time-variant parameters we choose Loan Age, Delinquency status, and Current Interest Rate, since our analysis showed these are the most important variables. The current interest rate always starts as the fixed interest rate, but can be updated over time. Incorporating these seven parameters results in the following results shown by Table 6.2. Since our computing power is not sufficient[4] to estimate parameters for the entire sample, we chose to take every 10^{th} observations into account. This means that we still take about 2.9 million observations into account.

First of all, we notice the AIC in Table 6.2, as explained in Appendix C, of 3,621,379 is much lower than the models we investigated in the Appendix F, indicating that this model performs better in terms of trade-off between goodness-of-fit and simplicity. Second, we notice that all coefficients are significant on a 99% level except for the LTV in state 3 and the LTV and DTI in state 2 and 3, which is not even significant on a 90% level. Hence, we do not take them into account.

Observing the constants it makes sense that all of them are negative, since this implies that being on Schedule has a higher basis chance compared to all other states. Looking at Table 6.1a we can explain that state 2 is most negative and state 4 is least negative considered that total number of observations present in these states as shown in Table 6.1a. Sometimes state 5 can have a counter intuitive sign, for example for delinquency status. This is due to the fact that after several months of delinquency mortgagors can choose to refinance or foreclose and make the full prepayment in order to get a new mortgage.

Since we are mainly interested in the risk of prepayment, we look for the differences in log odds ratio's for state 4. The constant tells us that, given all other variables are constant, we are 4.52 times more likely to be on schedule than in prepayment. The loan age coefficient of 0.013 tells us that, ceteris paribus, after 117 months the odds between being in state 1 and 4 are about equal. The odds of being in state 4, compared to state 1, are 14% lower in case a prepayment penalty is present. If the credit score is considered good, in other words 650, the odds of being in state 4 are multiplied by 3.7 implying

[4] 2.6 Ghz quad-core processor with 16GB RAM

that a score of higher than 755, the mortgagor is more likely to prepay than to stay on schedule. We observe that the higher the interest rate, the less likely the mortgagor is to prepay. This could possibly be explained by the fact that the mortgagor has insufficient funds to prepay due to the higher payments he has to make. Finally we note that the higher the DTI and LTV, the less likely the mortgagor is to prepay.

Economically the signs of these variables make sense. The longer the mortgage lasts, the more chance there is that the mortgagor will either prepay of get delinquent, hence all coefficients are positive. Delinquency status explains positive for delinquency and negative for prepayment. The higher the interest rate, the higher the chance of default and the lower the chance of prepayment. Higher credit scores imply more chance of prepayment and lower chance of delinquency and default. Higher LTV and DTI give a lower chance of prepayment due to the relative higher loan and debt which gives higher monthly cost and hence no funds to cover extra payment.

On top of this segmented model for five states the segmentation in state 4 is done, as shown in Equation (5.2) for different shares of prepayment. The results of this model can be found in Table 6.3. Note that we can compare the AIC of this model with the model above in Table 6.2 since both take the exact same 10% of observations into account. As we can see, the segmented model outperforms the subsegmented model in terms of AIC and therefore we prefer the model in Table 6.2. The segmented model in Table 6.3 shows that state 4.1-4.10 are the same in terms of sign and for 4.2-4.10 the values are also similar. Furthermore for state 4.2-4.10 current interest rate and FICO are monotonically increasing whereas Loan Age and DTI remains about constant. This tells us that Loan age and FICO are good estimators and not just a small peak for a few months prepayment.

Table 6.2: Results MNL Model 2001-2016 $\mathbb{P}$[On Schedule] ($k = 1$) vs. $\mathbb{P}$[Default], $\mathbb{P}$[Delinquent], $\mathbb{P}$[Part. Prepayment] and $\mathbb{P}$[Full Prepayment] ($\kappa = 2, \ldots, 5$) Loan Age, Delinquency Status, Interest Rate, FICO, LTV, DTI & Prepayment Penalty

	Dependent variable:			
	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
Loan Age	0.011***	0.012***	0.013***	0.005***
	(0.003)	(0.001)	(0.00005)	(0.0002)
Del Status	21.540***	13.790***	−0.034***	6.058***
	(0.0001)	(0.0001)	(0.00000)	(0.00003)
Cur Int Rate	0.011***	0.071***	−0.153***	0.228***
	(0.0001)	(0.0001)	(0.001)	(0.004)
FICO Score	−0.004***	−0.006***	0.002***	0.003***
	(0.001)	(0.0002)	(0.00001)	(0.00004)
LTV	0.010	−0.0004	−0.005***	−0.002***
	(0.006)	(0.001)	(0.0001)	(0.0003)
DTI	−0.012	0.006***	−0.019***	−0.005***
	(0.008)	(0.002)	(0.0001)	(0.0004)
Prepayment Penalty	0.957***	−0.436***	−0.151***	−0.522***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Constant (α)	−47.980***	−5.004***	−1.510***	−6.812***
	(0.00002)	(0.00001)	(0.00003)	(0.0001)
AIC	3,621,379			
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6.3: Results MNL Model 2001-2016 with sub-segmentation in κ=4 Loan Age, Current Int. Rate, FICO & DTI

	$\kappa = 2$	$\kappa = 3$	$\kappa = 4.1$	$\kappa = 4.2$	$\kappa = 4.3$	$\kappa = 4.4$	$\kappa = 4.5$
Loan Age	0.012***	0.012***	0.011***	0.026***	0.026***	0.025***	0.025***
	(0.001)	(0.0001)	(0.00005)	(0.0001)	(0.0002)	(0.0003)	(0.0003)
Cur Int	0.566***	0.391***	−0.142***	−0.321***	−0.318***	−0.295***	−0.244***
	(0.0001)	(0.004)	(0.001)	(0.004)	(0.006)	(0.008)	(0.010)
FICO	−0.010***	−0.013***	0.002***	0.007***	0.009***	0.010***	0.011***
	(0.0001)	(0.00003)	(0.00001)	(0.00003)	(0.00004)	(0.0001)	(0.0001)
DTI	0.026***	0.019***	−0.017***	−0.035***	−0.036***	−0.036***	−0.035***
	(0.002)	(0.0003)	(0.0001)	(0.0004)	(0.001)	(0.001)	(0.001)
Constant (α)	−4.504***	2.103***	−1.533***	−6.963***	−9.791***	−11.090***	−12.320***
	(0.00000)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)

	$\kappa = 4.6$	$\kappa = 4.7$	$\kappa = 4.8$	$\kappa = 4.9$	$\kappa = 4.10$	$\kappa = 5$
Loan Age	0.023***	0.024***	0.025***	0.025***	0.026***	0.005***
	(0.0003)	(0.0004)	(0.0004)	(0.0005)	(0.001)	(0.0002)
Cur Int	−0.197***	−0.180***	−0.166***	−0.147***	−0.162***	0.236***
	(0.011)	(0.013)	(0.015)	(0.0002)	(0.0001)	(0.004)
FICO	0.010***	0.011***	0.014***	0.013***	0.014***	0.003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00003)
DTI	−0.032***	−0.032***	−0.027***	−0.034***	−0.031***	−0.005***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.0004)
Constant (α)	−12.480***	−13.700***	−16.020***	−15.820***	−17.100***	−7.112***
	(0.0002)	(0.0003)	(0.0003)	(0.00000)	(0.00000)	(0.0001)
AIC:			5,289,387			
Note:						*p<0.1; **p<0.05; ***p<0.01

For the Survival Analysis we took the Fico score, Mortgage Insurance Percentage, Loan-to-Value, Combined LTV, Original Unpaid Principal Balance, Debt-to-Income, Interest Rate and Prepayment Penalty into account, since these are fixed variables that we can change in advance of closing the mortgage. The results of the survival analysis are shown in Table 6.4.

Table 6.4: Results Survival Analysis 2001-2016

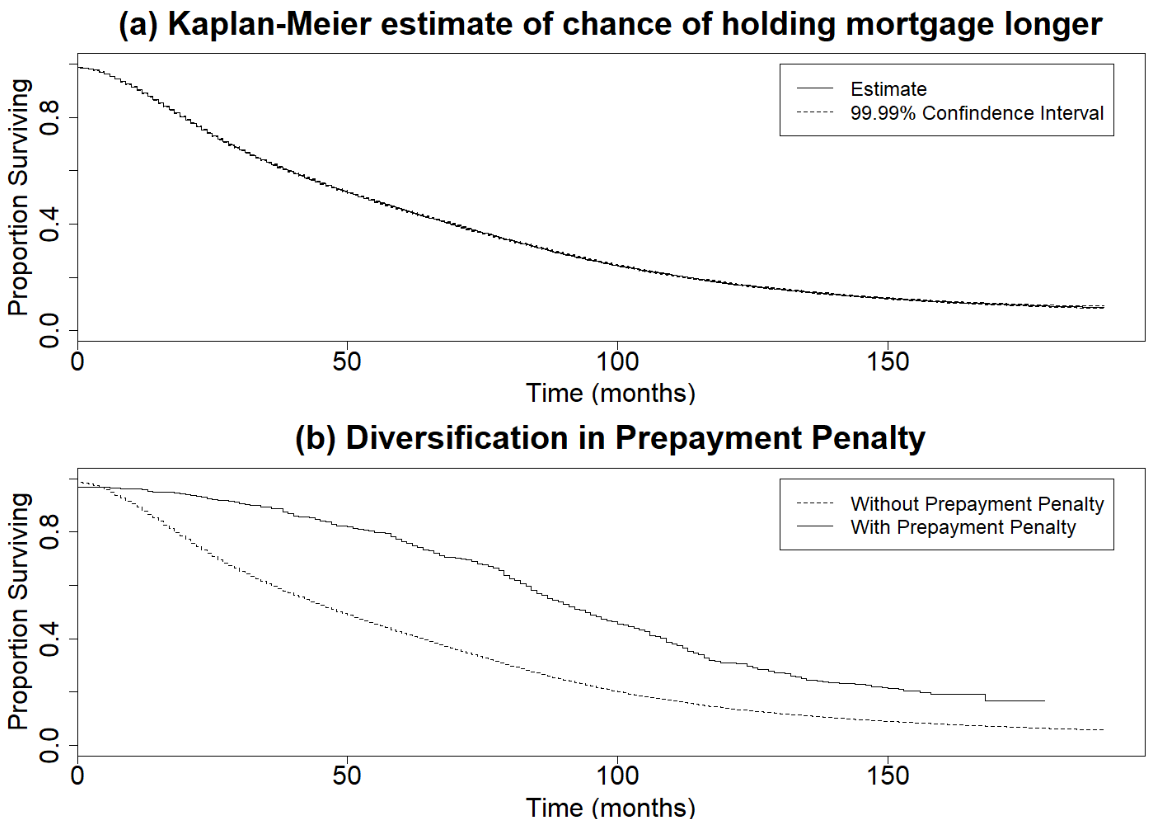
Variable k	*Coefficient* (β_k)
Fico Score	0.002^{***}
	(0.00003)
LTV	-0.004^{***}
	(0.0003)
CLTV	0.002^{***}
	(0.0003)
Original UPB	0.003^{***}
(× 1,000)	(0.00001)
DTI	-0.003^{***}
	(0.0001)
Interest Rate	0.404^{***}
	(0.002)
Prepayment	-0.799^{***}
Penalty	(0.046)
Observations	762,002
Pseudo R^2	0.101
Log-Likelihood	−6,300,421
Wald Test	$78{,}659^{***}$ (df = 7)
LR Test	$80{,}992^{***}$ (df = 7)
LM (score) Test	$79{,}092^{***}$ (df = 7)

Note: *p<0.1; **p<0.05; ***p<0.01

We observe that all effects are significant on a 99% level. The pseudo R^2 of (Cox and Snell, 1989) is 0.101, indicating that the model is an improvement of the model without explaining variables. Also, the Likelihood Ratio test shows that the model performs better than the benchmark model with no explaining parameters on a 99% level. The Wald test indicates that the parameters are satisfied and the LM test says that the log-likelihood of the benchmark model is not close enough to zero. Since "$Wald \leq LR \leq LM$" does not hold we conclude that the model is not linear, which makes sense given that we are dealing with a logistic regression. This models gave better results in terms of log-likelihood than the other models that have been tried. Further explanation about the R^2, Wald Test, LR and LM test can be found in Appendix C.

Taking a closer look at the coefficients, we conclude that Fico, CLTV, principal and interest rate have a negative effect on the lifetime of the mortgage, since increasing by one unit implies an increase of the hazard rate. An increase of LTV, DTI and Prepayment penalty results in a longer lifetime. We are not able to compare these values directly to the found coefficients in Section 6.1.1, but the interpretation is similar and there are no contradictions except for the interest rate. This difference will be discussed in the model comparison in Section 6.5. Note that some coefficients seem small at first, but that most units are considerably small as well. For example, an increased principal of €100,000 results in a mortgage that lasts 36% shorter. The Prepayment Penalty coefficient of -0.799 tells us that, ceteris paribus, mortgagors with a prepayment penalty clause are 55% less likely to end the mortgage. This makes sense, given the fact that they need to pay an additional amount in case they make a premature payment.

Figure 6.2: Survival Analysis Kaplan-Meier & Segmentation in Prepayment Penalty



As we can see clearly from Figure 6.2a the survival function follows a steep decent after the first few months with the lowest derivative (steepest descent) of the graph close to -1 around 20 months and recovers to a slow descent rate after 60 months. Furthermore we observe that the confidence interval, even though it is 99.99%, is very small, indicating that our estimate is very accurate. In Figure 6.2b we distinguished cases with the presence and absence of a prepayment penalty. It shows that mortgages with a prepayment penalty survive for a longer period of time. This makes sense, since prepaying is less attractive for these mortgagors.

6.3. Results Markov Model

For the same states as used in the MNL model we estimate the transition probabilities of the time-constant Markov Switching model. Since our goal is to investigate whether their exists a difference between transition probabilities in different times, three subsets are estimated. The first model contains transition probabilities from 2001-2005, the second ranges from 2006-2010, the third model incorporates 2011-2016 and the last models captures the entire set ranging from 2001-2016.

Tables 6.5-6.7 show the results of the Markov Switching models in the four different time periods for which state 1,...,5 correspond to the prepayment states as described in Equation (5.1). First of all, we note that all parameters that could be unequal to zero or one are significant on a 99% confidence interval. We can not compare the log-likelihoods directly to each other, since the time period and number of periods are simply not the same. However, we can say that despite the fact that Table 6.7 has more than half the size of observations that Tables 6.5 and 6.6 have, the log-likelihood is less than twice as low, indicating that 2011-2016 gives a better estimate than the two other periods. Besides that, the sum of log-likelihoods of Tables 6.5-6.7 is less than the log-likelihood of Table 6.8. This suggests that splitting our model in different time periods explains better.

The main insight is that the state in time $t + 1$ tends to stay at the same state as on time t for all states in all time periods. Only in case of being in state 3 the chance of going to state 1 (On schedule), π_{31}, is considerably big compared to all other transition probabilities. This could possibly be explained by the fact that mortgagors forgot to pay a month or simply could not afford the payment for some months. The fact that π_{31} is even higher in the time period 2011-2016 supports this theory, since this is the period after the crisis. Because π_{31} is higher than the probability to default, π_{32}, it indicates the average mortgagor has the willingness to pay the mortgage rather than to default.

Table 6.5: Markov Transition Probabilities Freddie Mac 2001-2005

(a) Estimates Markov Transition Probabilities Freddie Mac 2001-2005

	1	2	3	4	5
1	0.965*	0	0.216*	0.007*	0
2	0	1	0.031*	0	0
3	0.010*	0	0.689*	0.006*	0
4	0.008*	0	0.054*	0.970*	0
5	0.017*	0	0.010*	0.017*	1
LogLik:	-2,464,185		*p<0.01		

(b) Std. Error Markov Transition Probabilities Freddie Mac 2001-2005

	1	2	3	4	5
1	0.0003	0	0.001	0.00005	0
2	0	0.002	0.0003	0	0
3	0.00004	0	0.001	0.00004	0
4	0.00003	0	0.0004	0.001	0
5	0.00004	0	0.0002	0.0001	0.002
Observations:					14,598,681

Table 6.6: Markov Transition Probabilities Freddie Mac 2006-2010

(a) Estimates Markov Transition Probabilities Freddie Mac 2006-2010

	1	2	3	4	5
1	0.964*	0	0.168*	0.007*	0
2	0	1	0.046*	0	0
3	0.010*	0	0.744*	0.005*	0
4	0.009*	0	0.035*	0.967*	0
5	0.017*	0	0.006*	0.021*	1
LogLik:	-2,125,489		*p<0.01		

(b) Std. Error Markov Transition Probabilities Freddie Mac 2006-2010

	1	2	3	4	5
1	0.0004	0	0.001	0.0001	0
2	0	0.001	0.0004	0	0
3	0.00004	0	0.001	0.00004	0
4	0.00004	0	0.0003	0.001	0
5	0.00005	0	0.0001	0.0001	0.002
Observations:					12,672,339

Table 6.7: Markov Transition Probabilities Freddie Mac 2011-2016

(a) Estimates Markov Transition Probabilities Freddie Mac 2011-2016

	1	2	3	4	5
1	0.976*	0	0.330*	0.007*	0
2	0	1	0.017*	0	0
3	0.002*	0	0.572*	0.001*	0
4	0.011*	0	0.071*	0.981*	0
5	0.011*	0	0.011*	0.011*	1
LogLik:	-868,676		*p<0.01		

(b) Std. Error Markov Transition Probabilities Freddie Mac 2011-2016

	1	2	3	4	5
1	0.0005	0	0.003	0.0001	0
2	0	0.013	0.001	0	0
3	0.00002	0	0.005	0.00002	0
4	0.00005	0	0.002	0.001	0
5	0.00005	0	0.001	0.0001	0.004
Observations:					8,551,906

Table 6.8: Markov Transition Probabilities Freddie Mac 2001-2016

(a) Estimates Markov Transition Probabilities Freddie Mac 2001-2016

	1	2	3	4	5
1	0.967*	0	0.198*	0.007*	0
2	0	1	0.037*	0	0
3	0.008*	0	0.710*	0.004*	0
4	0.009*	0	0.046*	0.972*	0
5	0.015*	0	0.008*	0.017*	1
LogLik:				-5,495,990	

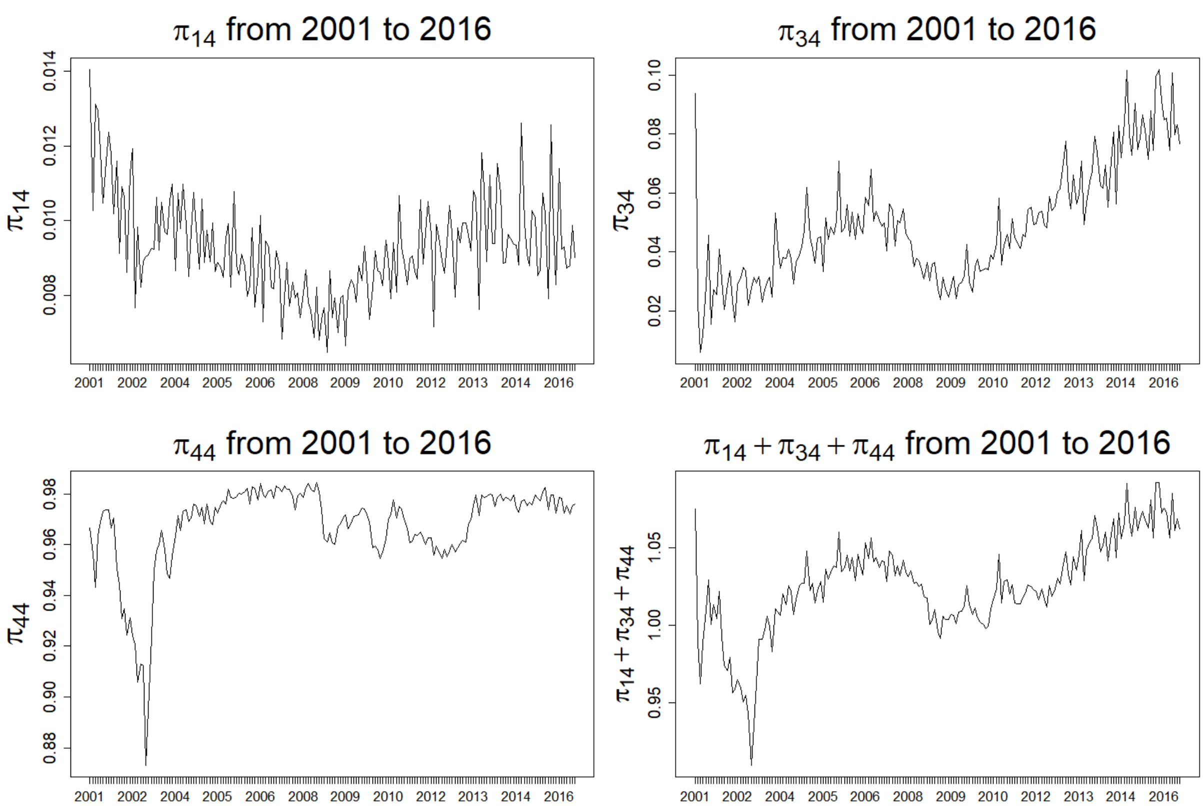
(b) Std. Error Markov Transition Probabilities Freddie Mac 2001-2016

	1	2	3	4	5
1	0.0002	0	0.001	0.00003	0
2	0.00000	0.001	0.0002	0.00000	0
3	0.00002	0	0.001	0.00002	0
4	0.00002	0	0.0003	0.0004	0
5	0.00003	0	0.0001	0.00005	0.001
Observations:					35,822,926

Since we are mainly interested in prepayment, we estimated a transition matrix for every single month and plotted the transition probabilities to state for, that is π_{i4} for i=1,3,4 in Figure 6.3. Al the other probabilities are plotted in Figures F.1-F.5 in Appendix F. From these Figures we observe that the proportion of prepayment is slightly increasing over time, with a little dip after 2008. However, the increase is small. The big dip in 2003 can be explained by the peak in full prepayments. This might be due to the relatively low interest at that time, as shown in Figure 2.1, which would impose a low mortgage rate. In that case refinancing would be a attractive alternative. If this is indeed the case and mortgagors would act similar, mortgagees could expect a refinance rate of about 5-10% of their mortgages.

Overall we can conclude that there exist different transition probabilities in different times, but that most of them are similar. None of the transition probabilities differ more than 0.05 from other time periods except for π_{3j}, especially π_{31} and π_{33}. Summing the rows, we see that moving to states 1 and 4 is bigger than zero for all periods versus state 3 being smaller than zero for all periods, indicating that mortgagors are more likely to stay on schedule or prepay instead of being delinquent.

Figure 6.3: Markov Transition Probabilities $\pi_{i4,t}$ for $i = 1, 3, 4$ and $t = 9, \ldots, 192$



6.4. Results Time-Varying Markov Switching Model

In order to estimate the time-varying Markov Switching model, we first estimate the time constant transition possibilities for every single month from 2001-2016. Since the first contract dates from February 2001 and the fact that we've deleted the first six observations, that is March 2001 until Augustus 2001, the first transition probabilities date from September 2001. The last transition matrix dates from December 2016. All together we are therefore left with 183 transition probability matrices. The results are plotted in Figure 6.3 above and Figures F.1-F.5 in Appendix F.

In order to estimate the model we took a burn-in period of 2 years, that is 24 observations ranging from September 2001 until Augustus 2013. Next we use the code of Bazzi et al. (2017) and rewrite this code from a 2 state model to a 5 state model. Note that since all probabilities add up to one, we need to estimate a $(K \times K)$ matrix for which we need to estimate only $K(K-1)$ probabilities. The same holds for the $\mathbf{f}_t, \omega, \mathbf{s}_t$, vectors in Equation (5.25). Since matrices $\mathbf{A}$ and $\mathbf{B}$ are assumed to be diagonal we also need to estimate $K(K-1)$ elements $(A_{11}, \ldots, A_{55}$ and $B_{11}, \ldots, B_{55})$ with one skipped estimate j on each A_{ij} and B_{ij}. It is easy to see that the multiplication factor c for an increase with a states compared to a two state model is:

$$c = \frac{(K+a)(K+a-1)}{K(K-1)} \overset{\text{K=2}}{=} \frac{(2+a)(1+a)}{2} = 1 + \frac{3a+a^2}{2}. \tag{6.2}$$

This implies that we need to estimate ten times as much values for each parameter, that is twenty instead of two. On top of that we need to estimate a Jacobian matrix of 400 (20×20) instead of four (2×2).

Since simply skipping every iK^{th} observation would lead to skewed estimates due to the multinomial logit specification in Equation (5.28) and the choice of $\delta_{ij} = 1e-10$ for all i, j, we choose to skip each estimate ψ_{ij} for which $j = i+1$ and $j = 1$ if $i = K$. Hence, in our 5 state model we skip $\psi_{12}, \psi_{23}, \psi_{34}, \psi_{45}$ and ψ_{51}.

Since we posses 16 years of data with monthly observations of 50,000 mortgagors per year, we estimated a transition probability matrix for every single month, leaving out the first seven observations plus the first (January 2001) and the last (December 2016) one since there was no data available for that month. In total we therefore hold 183 data points with 25 transition probabilities each. Eventually we are trying to estimate five means μ_i, one σ, 20 transition probabilities π_{ij} and in the most extensive GAS modelling framework 20 values for both A_{ij} and B_{ij}. Estimating the 66 parameters of this GAS model with only 183 observations leads to bad estimates, high standard errors and a non-singular Hessian Matrix due to the property of Rank-Deficiency. Therefore we interpolate the observed monthly data points to daily observations in the following way:

$$\hat{\pi}_{ij,t_k} \sim N\left(\frac{30-k}{30}\hat{\pi}_{ij,t} + \frac{k}{30}\hat{\pi}_{ij,t+1}, \quad \left[\frac{\hat{\sigma}_{ij,t} + \hat{\sigma}_{ij,t}}{2}\right]^2 \right), \qquad \forall\ k = 0, \ldots, 30. \quad (6.3)$$

Since the model is now also estimating probabilities that we already know, the following parameter are fixed to either one or zero:

$$\pi_{21} = \pi_{23} = \pi_{24} = \pi_{25} = 0, \quad \pi_{22} = 1, \qquad \pi_{51} = \pi_{52} = \pi_{53} = \pi_{54} = 0, \quad \pi_{55} = 1,$$
$$A_{21} = A_{22} = A_{23} = A_{24} = A_{25} = 0, \qquad A_{51} = A_{52} = A_{53} = A_{54} = A_{55} = 0,$$
$$B_{21} = B_{22} = B_{23} = B_{24} = B_{25} = 1, \qquad B_{51} = B_{52} = B_{53} = B_{54} = B_{55} = 1.$$

On top of that, in contradiction to Bazzi et al. (2017) we did not take alternative variances for different regimes into account, limiting our estimates even more. To take both dynamic parameters in account a difference is made between a GAS Model and TVP model. In case of the TVP the diagonal elements of **A** are estimated but **B** is fixed to an identity matrix. In case of GAS both the diagonal elements of **A** and **B** are estimated. Besides these time-varying models a static model (MS) is estimated. This leads to the results as shown below in Table 6.9.

Table 6.9: Results Constant Markov Switching model (MS), Time Varying Probabilities (TVP) and framework with generalized autoregressive score (GAS) including standard Error and T-test

$\boldsymbol{\theta}$	**Start**	**MS**	**SE**	**T-test**	**TVP**	**SE**	**T-test**	**GAS**	**SE**	**T-test**
μ_1	0	0.321	0.114	2.808	0.404	0.045	8.903	0.517	0.001	491.2
μ_2	-2	-2.020	15.6e3	0	-2	5.0e3	0	-0.504	0	3.7e3
μ_3	-1	-1.159	0.249	4.658	-1.075	0.129	8.317	-0.897	0.001	678.1
μ_4	1	0.245	0.030	8.169	0.250	0.029	8.737	0.227	0.001	350.8
μ_5	2	1.603	0.447	3.588	1.908	142.5	0.013	0.819	0.002	540.2
σ^2	0.500	1.191	0.074	16.06	1.293	0.041	31.76	1.320	0.002	764.0
π_{11}	0.967	0.632	0.116	5.455	0	0	96.55	0.617	0.003	197.3
π_{13}	0.008	0.263	0.011	23.08	0	0	18.68	0.379	0.006	60.55
π_{14}	0.009	0.002	0.001	1.342	1	0	678e3	0.005	0	79.02
π_{15}	0.015	0.104	0.059	1.766	0	0	0.424	0	0	0.075
π_{31}	0.198	0.792	0.335	2.363	0	0	1.014	0.111	0	$1,039$
π_{32}	0.037	0	0	0	0	0	0.005	0	0	0.003
π_{33}	0.710	0.208	0.315	0.660	1	0	26.9e6	0.887	0.001	$1,665$
π_{35}	0.008	0	0	0	0	0	0.012	0	0	0.063
π_{41}	0.007	0.001	0.011	0.051	0.001	0.014	0.081	0	0	459.2
π_{42}	0.001	0	0	0	0	0.014	0.003	0	0	0.187
π_{43}	0.004	0	0.001	0.011	0	0.037	0.005	0	0	4.424
π_{44}	0.971	0.998	0.012	84.78	0.997	0.012	79.89	1	0	3.6e5
A_{11}	0	-	-	-	-3.258	0.022	147.5	3.419	0.207	16.52
A_{13}	0	-	-	-	-3.920	0.037	106.2	0.471	40.74	0.012
A_{14}	0	-	-	-	-0.710	0.174	4.090	-0.032	15.67	0.002
A_{15}	0	-	-	-	-7.543	49e12	0	0.002	4.711	0
A_{31}	0	-	-	-	-23.05	28.58	0.806	3.410	0.040	84.78
A_{32}	0	-	-	-	-6.090	595e9	0	0.031	69.53	0
A_{33}	0	-	-	-	0.551	0.510	1.080	-0.050	0	851.7
A_{35}	0	-	-	-	-2.037	382.5	0.005	-0.006	4.097	0.002
A_{41}	0	-	-	-	2.067	3.460	0.597	1.551	0.006	277.6
A_{42}	0	-	-	-	-1.596	312.6	0.005	0.004	309e9	0
A_{43}	0	-	-	-	-1.104	59.27	0.019	-0.002	2.789	0.001
A_{44}	0	-	-	-	-0.753	0.523	1.439	-0.288	0.001	530.4
B_{11}	0.900	-	-	-	-	-	-	0.978	0.002	14.20
B_{13}	0.900	-	-	-	-	-	-	0.900	4.529	0.020
B_{14}	0.900	-	-	-	-	-	-	0.900	104.9	$9.54e{-4}$
B_{15}	0.900	-	-	-	-	-	-	0.900	76.34	$1.31e{-3}$

Continued on next page

Table 6.9: Results MS, TVP and GAS including SE and T-test

$\boldsymbol{\theta}$	**Start**	**MS**	**SE**	**T-test**	**TVP**	**SE**	**T-test**	**GAS**	**SE**	**T-test**
B_{31}	0.900	-	-	-	-	-	-	0.463	0.011	49.41
B_{32}	0.900	-	-	-	-	-	-	0.900	578.2	1.73e−4
B_{33}	0.900	-	-	-	-	-	-	0.983	0	267.7
B_{35}	0.900	-	-	-	-	-	-	0.900	226.7	4.41e−4
B_{41}	0.900	-	-	-	-	-	-	0.936	0	138.7
B_{42}	0.900	-	-	-	-	-	-	0.900	16e12	6.4e−15
B_{43}	0.900	-	-	-	-	-	-	0.900	428.5	2.33e−4
B_{44}	0.900	-	-	-	-	-	-	0.945	0	154.0
logLik	-	-5,906	-	-	-5,908	-	-	-5,893	-	-
AICc	-	11.8e3	-	-	11.9e3	-	-	11,87	-	-
BIC	-	11.9e3	-	-	12.0e3	-	-	12,05	-	-
MAE	-	0.953	-	-	0.952	-	-	0.948	-	-
MSE	-	1.499	-	-	1.496	-	-	1.482	-	-
MASE	-	0.779	-	-	0.780	-	-	0.778	-	-
MSSE	-	0.997	-	-	1.001	-	-	0.994	-	-

The first thing we notice is that some parameters go out of bound, like the standard error of μ_1 in the static (MS) and the TVP model. For the t-test a two-sided 95% confidence interval is used with an infinite degrees of freedom, that is a t-value of 1.96. Most parameters are significant, but the few that are not, are not significant at all. Despite the fact that we have enlarged our sample size, fixed 20% of parameters and assumed a regime independent variance, this results can most probably be explained by the fact that the model still has to estimate 30 parameters for the TVP model and 42 for the GAS model. Therefore, we drastically limit our 5-state model to a 2-state model. Similar to Equation (5.1) a segmentation over two states is made such that:

$$Y_{it} = \begin{cases} k = 1 : \text{On Schedule/Delinquent/Default}, & \text{if} \quad \text{Def}_i < \Delta_{it} \leq 0, \\ k = 2 : \text{Partial/Full Prepayment}, & \text{if} \quad 0 < \Delta_{it}. \end{cases} \tag{6.4}$$

The results of the 2-state model are shown below in Table 6.10. We observe that all parameters seem reasonable and within bound. On top op that, the standard errors seem small, which make all the parameters of the general static model significant. However, most values of **A** and **B** in the TVP and GAS model are not significantly different from zero or one on a 95% confidence interval. Taking a look at the overall performance, we observe that the GAS model outperforms both the static and TVP models in terms of log-likelihood. Furthermore, the TVP and GAS model both outperform the static model in terms of Bayesian Information Criterion (BIC) and AIC. The TVP loses in terms of log-likelihood to both other models.

Looking at the Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Scaled Error (MASE) and Mean Squared Scaled Error (MSSE), the time constant model always scores better than the TVP and performs better than or equal to the GAS model on all fronts. In case of the static model this can be explained since less parameters need to be estimated. However, the fact that TVP underperforms in terms of errors, indicates that only estimating $\mathbf{A}$ does not cover sufficient time-variation.

Table 6.10: Results 2-state model

$\boldsymbol{\theta}$	**Start**	**MS**	**SE**	**T-test**	**TVP**	**SE**	**T-test**	**GAS**	**SE**	**T-test**
μ_1	-1.000	0.209	0.044	4.765	-1.011	0.029	34.92	-1.008	0.029	35.30
μ_2	1.000	0.470	0.039	12.17	0.998	0.019	53.34	1.001	0.019	54.01
σ^2	0.500	1.371	0.032	42.42	0.498	0.016	31.63	0.497	0.016	31.76
π_{11}	0.972	0.996	0.003	367.6	0.314	0.022	14.31	0.314	0.026	11.90
π_{22}	0.988	0.997	0.002	416.3	0.665	0.017	39.36	0.676	0.029	23.07
A_{11}	0.000	-	-	-	-0.159	0.148	1.079	0.092	0.039	2.341
A_{22}	0.000	-	-	-	0.062	0.087	0.719	0.013	0.009	1.573
B_{11}	0.900	-	-	-	-	-	-	0.980	0.013	1.574
B_{22}	0.900	-	-	-	-	-	-	0.998	0.003	0.828
logLik	-	-5,784	-	-	-5,589	-	-	-5,579	-	-
AICc	-	11.6e3	-	-	11.2e3	-	-	11.2e3	-	-
BIC	-	11.6e3	-	-	11.2e3	-	-	11.2e3	-	-
MAE	-	0.976	-	-	0.979	-	-	0.976	-	-
MSE	-	1.381	-	-	1.387	-	-	1.383	-	-
MASE	-	0.830	-	-	0.831	-	-	0.830	-	-
MSSE	-	1.000	-	-	1.000	-	-	1.002	-	-

Deriving these results we impose that the model is not applicable in this case due to the insignificant parameters, which is probably stemming from the complexity of the model. Despite the fact that the time varying models score better in terms of AICc and BIC, they do not beat the static model in terms of MAE, MSE, MASE and MSSE. Besides that, the estimated parameters for $\boldsymbol{A}$ and $\boldsymbol{B}$ are not significantly different from zero and one respectively. Therefore we conclude that the time-varying model is not suited for this problem and we prefer the static model.

6.5. Model Comparison

First of all it needs to be stated that not all models can be compared directly in terms of log-likelihood, AIC, R^2 or any other criterium since they simply differ to much from each other. However, we are able to rate similar models like the two MNL models. Moreover, we can also compare the general performance, strong and weak points.

Overall we conclude that MNL model performs best, since it captures the effects accurately and the findings can economically be explained. On top of that, the lack of capturing the time dependency seems less important as shown by the individual Markov Switching models as well as the time varying transition probability model. The survival model gives insight in the factors that lead to the end of the mortgage. Since Table 4.1a tells us 96.08% of mortgage end is due to prepayment, we conclude that the factors the negatively influence the mortgage lifetime are indicators of prepayment.

All the results of the MNL and survival analysis are in accordance, except for the effect of interest rate. The MNL model tells us the likelihood of prepayment decreases when the interest rate increases, whereas the survival analysis shows us the lifetime and therefore the prepayment rate increases in case of higher interest rates. Since the result of the full prepayment in the MNL the likelihood of full prepayment increases, we conclude that increasing the interest rate leads to lower partial prepayments but a higher share of refinancing.

The Regular Markov Switching model shows us that the differences within different time intervals are small. However, a prepayment peak around 2003 is spotted which could indicate a increased risk for refinancing during low interest rate tides. The lack of time dependency gets confirmed by the time-varying Markov switching model, since the dynamic parameters are not significantly different from one and zero.

7. Conclusion and Discussion

7.1. Conclusion

In this research the risk of mortgage prepayment in the US housing market is investigated in order to check whether there exists a difference between different time periods. The reason for this research is the combination of rising housing prices and the lowering savings rate. Currently used models might not be able to capture the effects well and some motives might have been changed due to the dynamic macroeconomic environment.

As benchmark a MNL model, survival analysis and a Markov switching model are used. Since all of these methods fail to capture the element of time variation, a Markov switching model with time varying probabilities is built. For as far as we know a time-varying Markov switching model has not been used to model mortgage prepayment before. Except for the survival analysis, every model assumes five states for the presence of the mortgagor, that is: default, delinquency, on schedule, partial prepayment, full prepayment. This way it is possible to compute the relative likelihood of prepayment and the transition probabilities to prepayment over time for each mortgagor. From these states we observe that on average 26.84% of the mortgagors is in prepayment versus only 2.46% that is delinquent, indicating that prepayment risk is indeed a bigger risk than default risk.

The results of the MNL model tell us that for FICO scores higher than 755, the mortgagor is more likely to make a prepayment and the odds increase by 0.2% for every unit increase in the FICO score. Levying a prepayment penalty makes the mortgagor 14% less likely to make a prepayment. Furthermore increasing the interest rate, LTV and DTI leads to a decrease in prepayment, but may also cause an increased chance of delinquency. Taking a closer look at partial prepayment, we found that most prepayments are no more than 10% of the outstanding debt and that there are no big differences between subsets of prepayments.

From the survival analysis we learn that about 50% of the mortgagors hold their mortgage no longer than five years. The termination rate for the first years is higher than for the last years, indicating that mortgagors are more likely to hold the mortgage longer, given that they hold it for a longer time already. In combination with the findings in the MNL model that loan age increases the chance on prepayment, this is an important variable to take into account. Furthermore an increase in Fico, CLTV, original principal and interest rate has a negative effect on the lifetime of the mortgage, whereas LTV and

DTI have a positive effect.

The Markov model is estimated to give a brief indication whether the transitions probabilities are changing over time. Dividing the time sample in three subgroups shows that the differences are relatively small, and mortgagors tend to stay in the state they are currently in. However, we do observe that during the financial crisis in 2008, mortgagors had a harder time to come back from their delinquent payment state. To further investigate this time dependency within the transition probabilities, for each month the transition probability is estimated and plotted over time. Apart from the crisis, we observe that prepayment increases over time, although the proportion is still very small.

A time-varying Markov Switching model is applied to capture difference in time. This is done by dividing the estimated parameters in a static and a dynamic part, where a score of the predictive observations density is taken into account as well. We conclude that in case of five states the model was unable to estimate the parameters accurately, even when the number of observations is interpolated by a factor 30 and the least important parameters are fixed. The fact that the static model as well as the two dynamic models are unable to capture the time-variation is due to a combination of the absence of time-variation and the great number of parameters estimates, leading to inaccuracy. The two state model performs better than the five state model, but still has insignificant estimators.

Main findings of the research are that higher mortgage interest shrinks partial prepayment in the short term, but increases the likelihood of premature mortgage termination in the long term. The best estimators according to the MNL model are the loan age and the Fico. Sub-segmentation in partial prepayment does not lead to extra information. In addition, adding a prepayment penalty clause leads to 55% less premature termination, but makes the mortgage less attractive and can not always be levied.

Overall we conclude that we are not able to improve upon the currently used MNL model in terms of mortgage prepayment estimation. The prepayment risk shows little signs of time dependency and therefore low interest rate environments do not differ significantly from high interest rate environments. The five state model suffers from rank deficiency due to the large number of parameters. The two state model performs better in terms of log-likelihood but fails in terms of errors to the static competitor.

7.2. Discussion & Further research

This research focuses on the US residential market only. Using the same techniques on the European or commercial market could lead to alternative results because of the different stakes and interests. Also, the lack of the MNL model to deal with dependency between consecutive observations is a big discussion point, especially since this is a big feature of the mortgage data. Even though this dependency has been investigated by means of a Markov model, the knowledge is not incorporated in the estimates of the MNL model.

For further research we would discourage to try a three state GAS model, since this would only increase the parameter uncertainty. We do recommend to assume different variance for multiple regimes and a model with **A** fixed where only **B** gets estimated, since **A** shows higher standard errors. Another option which did not fit within the scope and time planning of this research is the latent variable analysis, as can be found in Appendix B.3. Because the prepayment rate did not seem to be time-dependent, there might be other latent variables that could explain the dynamics of the mortgage market better.

References

Akaike, H. (1974). “A new look at the statistical model identification”. *IEEE transactions on automatic control*, 19(6), pp. 716–723 (cited on p. 55).

Albert, J. H. and Chib, S. (1993). “Bayesian analysis of binary and polychotomous response data”. *Journal of the American statistical Association*, 88(422), pp. 669–679 (cited on p. 54).

Ambrose, B. W. and Sanders, A. B. (2003). “Commercial mortgage-backed securities: prepayment and default”. *The Journal of Real Estate Finance and Economics*, 26(2-3), pp. 179–196 (cited on p. 11).

Badarinza, C., Campbell, J. Y., and Ramadorai, T. (2017). “What calls to ARMs? International evidence on interest rates and the choice of adjustable-rate mortgages”. *Management Science* (cited on p. 7).

Bazzi, M., Blasques, F., Koopman, S. J., and Lucas, A. (2017). “Time-Varying Transition Probabilities for Markov Regime Switching Models”. *Journal of Time Series Analysis*, 38(3), pp. 458–478 (cited on pp. 1, 22–24, 36, 37).

Ben-Akiva, M., Bergman, M., Daly, A. J., and Ramaswamy, R. (1984). “Modeling interurban route choice behaviour”. In: *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*. VNU Science Press Utrecht, The Netherlands, pp. 299–330 (cited on p. 11).

Bhattacharya, A., Wilson, S. P., and Soyer, R. (2017). “A Bayesian approach to modeling mortgage default and prepayment”. *arXiv preprint arXiv: 1706.07677* (cited on p. 9).

Cavanaugh, J. E. (1997). “Unifying the derivations for the Akaike and corrected Akaike information criteria”. *Statistics & Probability Letters*, 33(2), pp. 201–208 (cited on p. 55).

Chernov, M., Dunn, B. R., and Longstaff, F. A. (2016). *Macroeconomic-driven prepayment risk and the valuation of mortgage-backed securities*. Tech. rep. National Bureau of Economic Research (cited on p. 1).

Consumer Financial Protection, B. of (2013). *Mortgage Servicing Rules Under the Real Estate Settlement Procedures Act (Regulation X), Doc.*

https://www.gpo.gov/fdsys/pkg/FR-2013-02-14/pdf/2013-01248.pdf (cited on p. 7).

Cox, D. R. (1992). “Regression models and life-tables”. In: *Breakthroughs in statistics*. Springer, pp. 527–541 (cited on p. 20).

Cox, D. R. and Snell, E. J. (1989). *Analysis of binary data*. Vol. 32. CRC Press (cited on pp. 31, 57).

Dekker, F. W., De Mutsert, R., Van Dijk, P. C., Zoccali, C., and Jager, K. J. (2008). “Survival analysis: time-dependent effects and time-varying risk factors”. *Kidney international*, 74(8), pp. 994–997 (cited on p. 10).

Deng, Y. and Liu, P. (2009). “Mortgage prepayment and default behavior with embedded forward contract risks in China’s housing market”. *The Journal of Real Estate Finance and Economics*, 38(3), pp. 214–240 (cited on p. 11).

Fisher, L. and Kan, J. (2015). “Chart of the week, Adjustable Rate Mortgage Share” (cited on p. 6).

Follain, J. R., Scott, L. O., and Yang, T. T. (1992). “Microfoundations of a mortgage prepayment function”. *The Journal of Real Estate Finance and Economics*, 5(2), pp. 197–217 (cited on p. 10).

FRED (2017). *Delinquency rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks*. URL: https://fred.stlouisfed.org/series/DRSFRMACBS (cited on p. 5).

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Vol. 138. CRC Press (cited on p. 55).

Heij, C., De Boer, P., Franses, P. H., Kloek, T., Van Dijk, H. K., et al. (2004). *Econometric methods with applications in business and economics*. OUP Oxford, pp. 230–243 (cited on pp. 57, 58).

Jaffee, D. (2003). “The interest rate risk of Fannie Mae and Freddie Mac”. *Journal of Financial Services Research*, 24(1), pp. 5–29 (cited on p. 9).

Kalotay, A., Yang, D., and Fabozzi, F. J. (2004). “An option-theoretic prepayment model for mortgages and mortgage-backed securities”. *International Journal of Theoretical and Applied Finance*, 7(08), pp. 949–978 (cited on p. 10).

Kelly, A. and Slawson, V. C. (2001). “Time-varying mortgage prepayment penalties”. *The Journal of Real Estate Finance and Economics*, 23(2), pp. 235–254 (cited on p. 10).

Lam, K., Schultz, J., and Raman, P. (2017). *The Size of the Affordable Mortgage Market: 2018-2020 Enterprise Single-Family Housing Goals*. Tech. rep. Federal Housing Finance Agency (cited on p. 7).

Liang, T.-H. and Lin, J.-B. (2014). “A two-stage segment and prediction model for mortgage prepayment prediction and management”. *International Journal of Forecasting*, 30(2), pp. 328–343 (cited on p. 11).

Märtens, K. and Ip, S. (2015). “Bayesian Logistic Regression with Polya-Gamma latent variables” (cited on p. 54).

Mattey, J. and Wallace, N. (2001). “Housing-price cycles and prepayment rates of US mortgage pools”. *The Journal of Real Estate Finance and Economics*, 23(2), pp. 161–184 (cited on p. 10).

Meis, J. (2015). “Modelling prepayment risk in residential mortgages” (cited on pp. 1, 10, 11, 18, 20).

Mills (2010). “Fundamentals on survival and event history analysis” (cited on p. 19).

Moench, E., Vickery, J. I., and Aragon, D. (2010). “Why is the market share of adjustable-rate mortgages so low?” (Cited on p. 7).

MotleyFool (2017). *30-Year vs. 5/1 ARM mortgage: Which Should I Pick?* URL: `https://www.fool.com/mortgages/2017/02/28/30-year-vs-51-arm-mortgage-which-should-i-pick.aspx` (cited on p. 7).

Netwerk, H. D. (2017). *Management Informatie*. URL: `https://www.hdn.nl/live/#management-informatie` (cited on p. 7).

Polson, N. G., Scott, J. G., and Windle, J. (2013). “Bayesian inference for logistic models using Pólya–Gamma latent variables”. *Journal of the American statistical Association*, 108(504), pp. 1339–1349 (cited on p. 54).

Popova, I., Popova, E., George, E. I., et al. (2008). “Bayesian forecasting of prepayment rates for individual pools of mortgages”. *Bayesian Analysis*, 3(2), pp. 393–426 (cited on p. 11).

Pravinvongvuth, S. and Chen, A. (2005). “Adaptation of the paired combinatorial logit model to the route choice problem”. *Transportmetrica*, 1(3), pp. 223–240 (cited on p. 10).

Quercia, R. G., Pennington-Cross, A., and Yue Tian, C. (2012). “Mortgage Default and Prepayment Risks among Moderate-and Low-Income Households”. *Real Estate Economics*, 40, S159–S198 (cited on p. 11).

Reserve, F. (2017). *Mortgage Debt outstanding*. URL: https://www.federalreserve.gov/data/mortoutstand/current.htm (cited on pp. 1, 5).

Rodrıguez, G. (2005). “Non-parametric estimation in survival models” (cited on p. 20).

Schwartz, E. S. and Torous, W. N. (1989). “Prepayment and the Valuation of Mortgage-Backed Securities”. *The Journal of Finance*, 44(2), pp. 375–392 (cited on pp. 1, 20).

— (1993). “Mortgage prepayment and default decisions: A Poisson regression approach”. *Real Estate Economics*, 21(4), pp. 431–449 (cited on p. 11).

Sheffi, Y. (1985). *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. ”” Prentice Hall, England Cliffs (cited on p. 10).

Vasconcelos, P. (2010). “Modelling Prepayment Risk: Multinomial Logit Model Approach For Assessing Conditional Prepayment Rate”. MA thesis. University of Twente (cited on pp. 1, 18).

Wit, E., Heuvel, E. v. d., and Romeijn, J.-W. (2012). “‘All models are wrong...’: an introduction to model uncertainty”. *Statistica Neerlandica*, 66(3), pp. 217–236 (cited on p. 55).

Zillow (2018). *United States home prices and values*. URL: https://www.zillow.com/home-values/#metric=mt%5C%3D4%5C%26dt%5C%3D1%5C%26tp%5C%3D5%5C%26rt%5C%3D14%5C%26r%5C%3D102001%5C%2C394913%5C%2C394806%5C%2C394463 (cited on p. 6).

Appendices

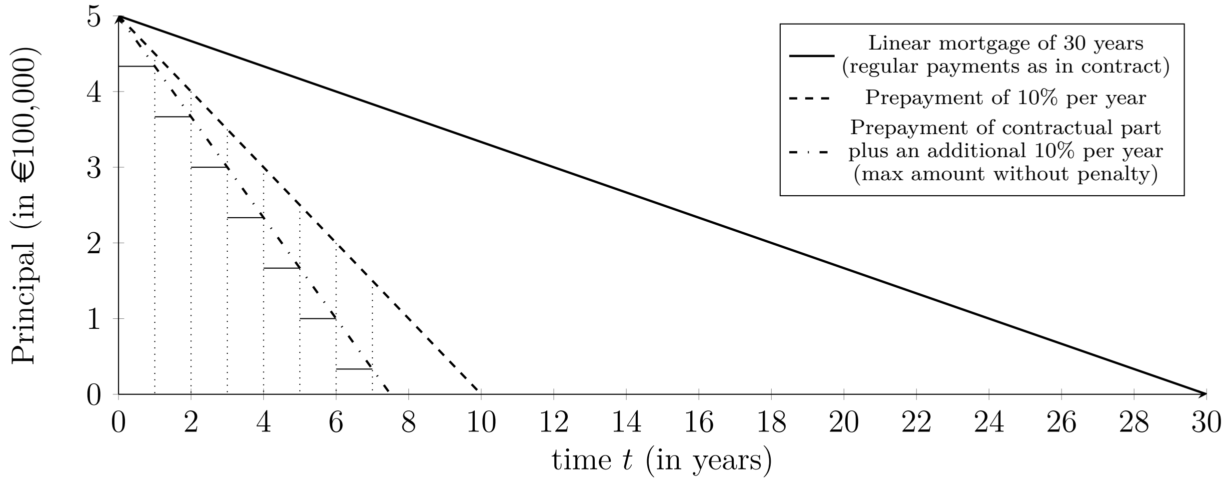
A. Prepayment Penalties

As stated in Section 1.2 and 2.1, prepayment penalties are often used in the European market to cover a part of the prepayment risk. However, they do not cover all risk, even if they fully apply. To show that raising prepayment penalties are no solid solution for the problem, an example of the situation is given.

In Figure A.1 we can find an example of a linear 30 year mortgage, which is given by the solid line. We assume that this mortgage has a clause for which the mortgagor can prepay a maximum of 10% of the original principal without paying a penalty on top of the contractual monthly payments. The dashed line indicates the situation where we make payments of 10% per year. The dash-dotted line therefore indicates the linear maximum 'speed' to which this mortgage can be payed off without paying a penalty, whereas the horizontal lines indicates the yearly bottoms and hence the maximum amount that can be payed off without making additional cost. The dotted vertical lines indicate the begin of every new year, which means that the mortgagor is allowed to make another prepayment of 10% of the principal again without paying a penalty. From the perspective of the mortgagee, the worst case scenario is when the mortgagor follows the black lines every year from the start of every new year, since in that case the mortgagor pays the least interest without paying any penalty to compensate for this lack of paid interest.

The expected loss can be calculated and explained by means of a simple example. If we assume a principal of €500,000, a loan term of 30 years and a yearly interest rate of 4% (0.33% per month in case of nominal compounding), under contractual payment the mortgagee receives a total amount of €359,246.40 on interest. However, if the mortgagor decides to payoff according to the green line with black bottoms in Figure A.1, the mortgagee only receives a total amount of interest of €70,315.70. In other words, the mortgagee loses a scheduled income of €288,930.70, which is more than 80% of his expected income, even if the prepayment penalties fully apply.

Figure A.1: Prepayment without penalty (client perspective)



B. Possible extensions on indirect effects

B.1. Indirect effects

In case of variables that explain the amount of prepayment in different ways than a direct effect, we take a look at Figure B.1, which distinguishes the four different types of effects. In Figure B.1a X_1 has a direct effect on Y. The size and sometimes sign of the correlation depends on modulator variable X_2. In case of an economic climate with low saving rates, keeping your savings at the bank is less attractive, making alternatives more likely.

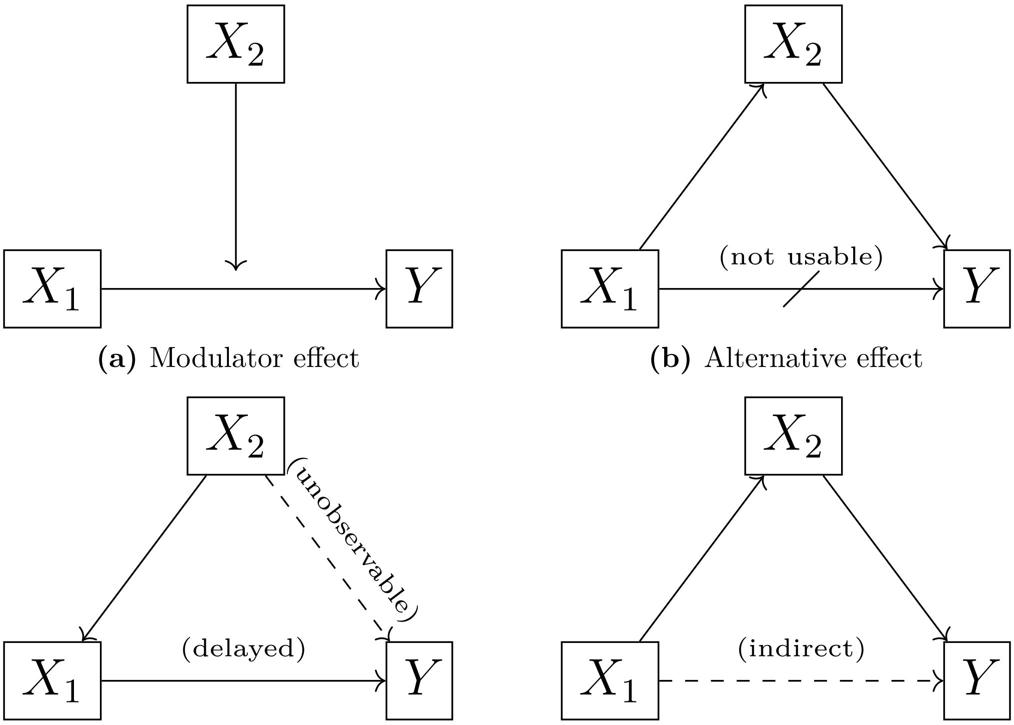
In Figure B.1b we are interested in the effect that X_1 has an on Y, but we are not allowed to use this variable because of privacy or discrimination policy for example. However, we are able to find a variable X_2 for which there is a correlation with X_1 and that has no such issues.

In Figure B.1c we know that X_2 has an effect on Y, but we do not observe this effect, since variable X_2 is not available in our database. However, we do posses a variable X_1 for which we know it is also influenced by X_2 and hence we are still able to observe it's effect on Y. Since this is an indirect effect, it might be the delayed effect of X_2 on Y.

Figure B.1d is very similar to Figure B.1c, but just the other way around. We observe an effect from X_1 on Y, for which we know that this is an indirect effect, since X_1 influences X_2 which influences Y on it's turn. Using X_1 we are able to estimate the effect of X_2. Note that this effect is similar to the effect in B.1b. The difference is that for the alternative effect we are not allowed to use the direct effect and therefore use the alternative effect. For the indirect effect we do use the direct effect, knowing that there no direct relation between X_1 and Y.

In this research we choose to mainly focus on the modulator effect as shown in Figure B.1a. More possible values for X_1 and X_2 in Figure B.1a and B.1b can be found below in Tables B.2 and B.3.

Figure B.1: Explanatory Models



(a) Modulator effect **(b)** Alternative effect

(c) Delayed effect **(d)** Indirect effect

B.2. Homogeneous groups

Another way to model the effects might be to make a segmentation in the variables by dividing them into subgroups to check whether there are different effects for different values of a certain variable. Below in Table B.1 some possibilities are given.

Table B.1: Homogeneous groups

Variable X_1	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**
Age (years)	0-25	26-35	36-45	46-70	>71
Income (€, annually)	0-25k*	25-40k	40-100k	100-500k	>500k
Zip Code	AA-EE	FF-JJ	KK-OO	PP-TT	UU-ZZ
Loan Age (years)	0-2	2-4	4-6	6-10	>10
...		...			

*k indicates $\times 1,000$, **m indicates $\times 1,000,000$

Table B.2: Modulator effects as in Figure B.1a

Variable X_1	**Variable X_2**	**How (Medium)**
Income	Expenditure Saving interest rate Inflation	
Mortgage Interest rate	Saving Interest rate Market mortgage rate Loan size	
Mortgage house price	Housing prices	

Table B.3: Indirect Factors as in Figure B.1b

Variable X_1	**Variable X_2**	**How (Medium)**
Income in t years	Bank Account Value	
Loan Age	Mortgagor Age	
Ethnic Background	Zip code	

Table B.4: Latent Variables

Latent variable	**Usefull macro economic variable***	
Faith in Housing market	Housing prices Inflation	
	Number of new house build Number of house demolished	Trend number of houses
Willingness to buy a house	Number of mortgages Number of people	Percentage of people buying a house
	Number of houses for rent Number of houses for sale Number of house seekers	
Believe about the savings rate	Investors long/short in saving rate	

* The macro economic variables can be obtained from Bloomberg

Table B.5: Unobserved variables and possible scraping data to forecast

Unobservable Variable	**Scraping data**	**Medium**
Birth of child	Pictures/Messages	Facebook/Instagram
	Payment information	Babyshop
Purchase Car/Boat/etc..	Payment Data	Garage/SHP
Marriage	Pictures/Messages	Facebook/Instagram
	Purchase ring/dress	Jeweller/Marriage shop
Bonus/Raise	Status update	LinkedIn
	Visiting hospital/Retirement	Hospital
Heritage/Donation*	Number of siblings	Facebook
	Intention	Funda
Move	Place of work	LinkedIn
	Place of payments	Bank Account
Holiday	Travel history	Facebook / Travel Blog
(Student) Loan*	Value of loan	Own dataset
	Interest rate	DUO / own dataset
Expenditure**	Trend expenditure	Bank account
(Private) Equity**	Savings account	Bank account
	Asset/Stock portfolio	Broker/Bank Account
Equity Parents**	Savings account	Bank account
	Asset/Stock portfolio	Broker/Bank Account
...	...	...

* Very hard   ** Only realistic if savings and mortgage at the same bank

B.3. Bayesian Logistic Regression using Pólya-Gamma Latent Variables

It is possible to model latent variables by means of a bayesian technique that is not often used. Our goal here is to sample from the posterior distribution of $\boldsymbol{\beta}$. In order to do so, we introduce the Pólya-Gamma distribution, which is distributed in the following way:

$$X \sim PG(b, c), \qquad x = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - \frac{1}{2})^2 + \frac{c^2}{4\pi^2}},$$

$$g_k \sim \text{i.i.d. } Gamma(b, 1), \qquad \text{Independent Gamma distributions } \forall\ k.$$

We assume a binomial likelihood of y_i, given $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$, and a Gaussian prior distribution of $\boldsymbol{\beta}$, such that

$$\text{Likelihood: } y_i | \boldsymbol{x_i}, \boldsymbol{\beta} \sim Bin\left(n_i, \frac{1}{1 + e^{-\psi_i}}\right),$$

$$\text{Prior: } \boldsymbol{\beta} \sim N(\boldsymbol{b}, \boldsymbol{B}).$$

Since we want to sample from the posterior of $\boldsymbol{\beta}$, we use a Pólya-Gamma latent variable for the following form:

Pólya-Gamma latent variable: $\omega_i|\boldsymbol{\beta} \sim PG(n_i, \boldsymbol{x}_i'\boldsymbol{\beta})$,

Posterior: $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\omega} \sim N(\boldsymbol{m_\omega}, \boldsymbol{V_\omega})$,

with mean and variance:

$$\boldsymbol{m_\omega} = \boldsymbol{V_\omega}(\boldsymbol{X}'\boldsymbol{k} + \boldsymbol{B}^{-1}\boldsymbol{b}), \quad \text{for} \quad \boldsymbol{k} = (y_1 - \frac{n_1}{2}, \ldots, y_N - \frac{n_N}{2}),$$

$$\boldsymbol{V_\omega} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{B}^{-1})^{-1}, \quad \text{where} \quad \boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega}).$$

Polson et al. (2013) and Märtens and Ip (2015) show that in this case the likelihood is given by:

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\omega} \sim N((\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}(\boldsymbol{X}'k), (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}),$$

where y_i = the number of successes, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})$ = vector of regressors for observation $i = 1, \ldots, N$, $\psi_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ = log odds of successes and n_i = number of trials.

Polson et al. (2013) state in their paper that Pólya-Gamma performs very well and only loses to the Metropolis Hastings sampler in case of logit models with abundant data and no hierarchical structure. But even here Pólya-Gamma is a close second. In our case we do posses abundant data, but since there is hierarchical structure present in the data, such as the Fico score and DTI, we might improve the model. The main differences compared to the paper of Albert and Chib (1993) are that the posterior is now a scale mixture instead of a location mixture of Guassians and that truncated normals are replaced by Pólya-Gamma latent variables.

C. Model fit, Error estimators & Statistical tests

C.1. Bayesian Information Criterion

According to (Wit et al., 2012) the BIC is defined by:

$$\text{BIC} = \kappa \ln(n) - 2 \ln(\hat{L}), \tag{C.1}$$

where n is the sample size, κ the number of parameters (θ) that need to be estimated by the model, and $\hat{L}$ the maximized value of the likelihood function calculated by using the maximum likelihood values ($\hat{\theta}$) for θ. Besides the fact that this information criterion is only valid for a sample size n larger than the estimated parameters κ, Giraud (2014) also argues that the BIC faces difficulties when exposed to high-dimensional problems.

C.2. Akaike Information Criterion

Akaike (1974) set up another criterion to measure the performance of a model. It can be used for model selection and makes a trade-off between the goodness-of-fit and the simplicity of the model. Since the AIC runs the risk of overfitting (Cavanaugh, 1997), AICc is founded and is basically AIC plus a correction for small sample sizes. The AIC and AICc are stated by:

$$\text{AIC} = 2\kappa - 2\ln(\hat{L}), \qquad \text{AICc} = \text{AIC} + \frac{2\kappa^2 + 2\kappa}{n - \kappa - 1}, \tag{C.2, 3}$$

where n, κ and $\hat{L}$ represent the same values as with the BIC.

Compared to the BIC the AIC punishes the number of parameters in a different way. With BIC the penalty is $\kappa \ln(n)$ whereas with AIC the penalty is equal to 2κ. The AICc has an additional punishment for the number of parameters

C.3. Mean Absolute Error

The MAE is given by:

$$\text{MAE} = \frac{1}{T - B + 1} \sum_{t=B}^{T} \left| y_t - \sum_{k=1}^{K} \xi_{t-1,k} \theta_k \right|, \tag{C.4}$$

where T is the sample size, B the burn in period so that we can see $T - B + 1$ as the size of our used sample. K is the number of state parameters, y and ξ are paired observations as explained in Equations (5.23) and (5.24) and $\theta_1, \ldots, \theta_K$ the state parameters that represent $\mu_1, \ldots, \mu_K$ in our case.

C.4. Mean Squared Error

The MSE is given by

$$\mathrm{MSE} = \frac{1}{T-B+1}\sum_{t=B}^{T}(y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k)^2, \tag{C.5}$$

where the parameters represent the same values as for the MAE. In contradiction to the MAE, the MSE gives more weight to higher errors because of the quadratic term.

C.5. Mean Absolute Scaled Error

The MASE is given by

$$\mathrm{MASE} = \frac{1}{T-B+1}\sum_{t=B}^{T}\left|\frac{\sum_{t=B}^{T} y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k}{\sqrt{\sigma^2 + \sum_{k=1}^{K}\xi_{t-1,k}(\theta_k - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k)^2}}\right| \tag{C.6}$$

$$= \cdots = \frac{\frac{1}{T-B+1}\sum_{t=B}^{T}\left|y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k\right|}{\frac{1}{T-B}\sum_{t=B+1}^{T}\left|y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k\right|} = \frac{\mathrm{MAE}}{\mathrm{MAE}^*}, \tag{C.7}$$

Compared to the MAE the MASE is scaled by the naive benchmark forecast given by MAE^*.

C.6. Mean Squared Scaled Error

The MSSE is given by

$$\mathrm{MSSE} = \frac{1}{T-B+1}\sum_{t=B}^{T}\left(\frac{\sum_{t=B}^{T} y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k}{\sqrt{\sigma^2 + \sum_{k=1}^{K}\xi_{t-1,k}(\theta_k - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k)^2}}\right)^2 \tag{C.8}$$

$$= \cdots = \frac{\frac{1}{T-B+1}\sum_{t=B}^{T}\left(y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k\right)^2}{\frac{1}{T-B}\sum_{t=B+1}^{T}\left(y_t - \sum_{k=1}^{K}\xi_{t-1,k}\theta_k\right)^2} = \frac{\mathrm{MSE}}{\mathrm{MSE}^*}, \tag{C.9}$$

Same as for the MASE the MSSE represents the MSE scaled by the naive benchmark forecast given by MSE^*.

C.7. Cox & Snell Pseudo R^2

The regular R^2, where y_i are the observations $i = 1, \ldots, n$ ranges between 0 and 1. 0 indicates a model with no explaining power and 1 indicates that the model explains perfectly.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}. \tag{C.10}$$

In contradiction to the regular R^2 often used in OLS regressions, the pseudo R^2 of Cox and Snell (1989) compares the likelihood of the model with no explanatory variables to the alternative model with several explanatory variables.

$$R^2_{cs} = 1 - \left(\frac{2L(M_0)}{2L(M_a)}\right)^{\frac{2}{n}}, \tag{C.11}$$

where n is the number of observations, L(.) equals the Likelihood and M_0,M_a represent the benchmark model and alternative model respectively. Theoretically the value of this R^2 can not reach the value of one, since in case of a perfectly explaining alternative model the upper bound is equal to $1 - L(M_0)^{\frac{n}{2}} < 1$.

C.8. Likelihood Ratio Test

The Likelihood Ratio test captures the loss of log-likelihood that stems from the parameter restrictions and therefore compares the current models with a benchmark model. The test statistic is calculated as follows:

$$LR = -2 \ln \frac{L(M_0)}{L(M_a)} \xrightarrow{d} \chi^2(g), \tag{C.12}$$

with g the degrees of freedom. Statistically this is the test with the most power. However, to compute this test we need to estimate 2 models.

C.9. Wald test

In contradiction to the LR test, the Wald test is only based on the restricted model with parameter estimates and tests tot what extend these parameter restrictions are satisfied by unrestricted estimates $\hat{\theta}_1$. According to Heij et al. (2004), under null hypothesis we have that:

$$W = r(\hat{\theta}_1)'(R_1 \mathcal{I}_n^{-1}(\hat{\theta}_1)R_1')^{-1} r(\hat{\theta}_1) \xrightarrow{d} \chi^2(g), \tag{C.13}$$

where $R_1 = \partial/\partial\theta'$ evaluated at $\theta = \hat{\theta}_1$, $\mathcal{I}_n$ the information matrix for sample size n and $r(.)$ the restrictions. This test performs well in case of difficult models such as non-linear parameter restrictions, but also depends on the parametization.

C.10. Lagrange Multiplier (Score) test

The LM test measures whether the gradient is close enough to zero at the restricted parameter $\hat{\theta}_0$. According to Heij et al. (2004) the test statistic is computed as follows:

$$LM = \left(\frac{\partial \log L(\theta)}{\partial \theta}\right)' \left(-E\left[\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'}\right]\right)^{-1} \left(\frac{\partial \log L(\theta)}{\partial \theta}\right) \xrightarrow{d} \chi^2(g). \tag{C.14}$$

An advantage of this test is that it requires simple computations, but the downfall is that the power of the test may be small.

D. Data description

Table D.1: Variable Description Freddie Mac Database Origination File

Variable	**Description (unit)**
Credit score (fico)	A score, prepared by third parties, to summarise the borrower's creditworthiness. This scores varies between 301-850. The higher the score, the better.
First Payment Date (dt_first_pi)	Date of the first scheduled mortgage note (YYYYMM).
First Home Flag (flag_fthb)	Indicates whether this is the first house bought. Takes values 1 (first house) and 0 (not the first house).
Maturity Date (dt_matr)	The month the last payment is made according to contract (YYYYMM).
MSA (cd_msa)	Metropolitan Division Metropolitan Statistical Area code.
Mortgage Insurance % (mi_pct)	Percentage of loss coverage on the loan in case of default (between 0-55%).
Number of Units (cnt_units)	Denotes if mortgage is a one-, two-, three- or four-unit property.
Occupancy Status (occpy_sts)	Denotes whether the mortgage type is owner occupied (1), second home (2) or Investment Property (3).
Combined Loan-to-Value (cltv)	In some cases there is a second loan. We add this other loan to our original loan and divide this by the value of the underlying property to obtain the CLTV (Between 0-200%).
Debt-to-Income (dti)	Monthly debt payments divided by monthly income (0-65%).
Unpaid Principal Balance (orig_upb)	UPB of the mortgage on note date ($).
Loan-to-Value (ltv)	Original loan amount divided by value of the underlying property(6-105%)
Interest Rate (int_rt)	Interest rate as on mortgage note (%).
Channel (channel)	Disclosure indicates the involvement of a Third party. Retail (1), Broker (2), Correspondent (3) or not specified (4).
Prepayment Penalty Flag (ppmt_pnlty)	Indicates there is a Prepayment Penalty (1) or not (0).
Product type (prod_type)	Denotes that the product is a Fixed Rate Mortgage (FRM).
Property State (st)	A two letter abbreviation for the state or territory (AL, TX, VA, etc.).
Property Type (prop_type)	Denotes whether the Property is secured by a Condominium (1), Leasehold (2), Planned Unit Development (3), Cooperative share (4), Manufactured Home (5) or Single Family Home (6).
Postal Code (zipcode)	Postal code (###00).

Continued on next page

Table D.1: Variable Description Freddie Mac Database Origination File (Continued)

Variable (Short)	**Description (unit)**
Loan Sequence Number (id_loan)	Unique ID of the loan (F1YYQnXXXXXX).
Loan Purpose (loan_purpose)	Indicates whether the loan is a Purchase (1), Cash-out Refinance (2), or No-Cash-out Refinance (3) mortgage.
Loan Term (orig_loan_term)	Number of scheduled months until Maturity (months).
Number of borrowers (cnt_borr)	Number of borrowers (1 or 2).
Seller name (seller_name)	Entity acting as the seller of the mortgage to Freddie Mac.
Servicer name (servicer_name)	Entity acting as the servicer of the mortgage to Freddie Mac.
Super Conforming Flag (flag_sc)	Indicates if mortgage exceeds conforming loan limits (1=Yes, 0=No).

Table D.2: Variable Description Freddie Mac Database Monthly Performance file

Variable	**Description (unit)**
Loan Sequence Number (id_loan)	Unique ID of the loan (F1YYQnXXXXXX).
Monthly Reporting Period (svcg_cycle)	The as-of month for loan information contained in the loan record (YYYYMM).
Current Actual UPB (current_upb)	Interest bearing UPB + non-interest bearing UPB($).
Loan Delinquency Status (delq_sts)	Indicates the delay in days.. 0-29 (1), 30-59 (2),..., or REO (-1).
Loan age (loan_age)	The number of months since the note origination month of the mortgage (months).
Remaining months to maturity (mths_remng)	Number of months to maturity (months).
Repurchase flag (repch_flag)	Indicates whether the loan is repurchased (1=Yes, 0=No).
Modification flag(flag_mod)	Indicates whether the loan is modified (1=Yes, 0=No).
Zero Balance Code (cd_zero_bal)	Indicates the reason why the loan is reduced to zero: Voluntary(1), Foreclosed by Alternative Group(2), Repurchase prior to Disposition(3) or REO Disposition(4).
Zero Balance effective date (dt_zero_bal)	Date on which the event took place (YYYYMM).
Current Interest rate (current_int_rt)	Denotes the current interest rate on the mortgage note (%).
Current deferred UPB (non_int_brng_upb)	Current non-interest bearing UPB of the modified mortgage ($).
Due date last paid Installment (dt_lst_pi)	Due date scheduled principal and interest is paid. (YYYYMM).

Continued on next page

Table D.2: Variable Description Freddie Mac Database Monthly Performance file (Cont.)

Variable (Short)	**Description (unit)**
MI Recoveries (mi_recoveries)	Mortgage Insurance (MI) recoveries: Proceeds received by Freddie Mac in the event of credit loss ($).
Net Sales Proceeds (net_sales_proceeds)	The amount remitted to Freddie Mac resulting from a property disposition ($). Covered (1) or Uncovered (0).
Non MI Recoveries (non_mi_recoveries)	Proceeds received by Freddie Mac based on repurchase/make whole proceeds, non-sale income such as refunds (tax or insurance), hazard insurance proceeds, rental receipts, positive escrow and/or other miscellaneous credits ($).
Expenses (expenses)	Expenses made by Freddie Mac in acquiring maintaining and/or disposing a property ($).
Legal Cost (legal_costs)	Amount of legal cost associated with the sale of the property ($).
Maintenance & Preservation Costs (maint_pres_costs)	Maintenance & Preservation costs associated with the sale of the property ($).
Taxes and Insurance (taxes_ins_costs)	Amount of taxes and insurance owed that are associated with the sale of the property ($).
Miscellaneous Expenses (misc_costs)	Miscellaneous Expenses associated with the sale of a property ($).
Actual Loss (actual_loss)	Actual loss = (Default UPB - Net Sale Proceeds) + Delinquent Accrued Interest - Expenses - MI Recoveries - Non MI Recoveries ($).
Modification Cost (modcost)	The cumulative modification cost amount calculated when Freddie Mac determines such mortgage loan has experienced a rate modification event ($).

E. Summary statistics

In this Appendix summary statistics of the origination file and the svcgfile are given. The origination file consists of variables that are fixed over time and are known before the closure of a mortgage. The svcgfile consists of monthly observation, meaning that for every single mortgagor from the origfile it can contain up to 360 observations. For each of the files the number of observations (N), the mean, the standard deviation, the minimum and the maximum is given. Table E.1 contains the summary of the origination file, whereas Table E.2 contains the summary stats of the svcgfile.

Taking a look at Table E.1 we observe that most of the values are present for every variable. The only variable that is missing about 34.8% of values is the first home indicator. Since this is a variable that that is not used in the analysis, we simply ignore the missing values. We observe that most credit scores (FICO) are far above 650 and therefore considered safe. The average interest rate is about 5.3% and the height of the mortgage differs a lot.

Table E.1: Summary statistics Origination File (origfile) Sample 2001-2016

Statistic	N	Mean	St. Dev.	Min	Max
fico	774,321	740.27	53.23	300	850
dt_first_pi	775,000	200,848.6	448.26	200,102	201,707
flag_fthb	505,344	0.17	0.372	0	1
dt_matr	775,000	203,569.3	623.25	201,003	205,704
mi_pct	774,855	4.44	9.954	0	50
cnt_units	774,993	1.03	0.220	1	4
occpy_sts	775,000	1.14	0.454	1	3
cltv	774,970	72.33	17.18	6	181
dti	765,853	33.70	11.27	1	65
orig_upb	775,000	194,462.3	107,454.3	8,000	1,144,000
ltv	774,975	71.13	16.90	6	101
int_rt	775,000	5.26	1.194	2.250	11.490
ppmt_pnlty	771,824	0.001	0.029	0	1
loan_purpose	775,000	1.93	0.845	1	3
orig_loan_term	775,000	33.41	68.73	60	604
cnt_borr	774,877	1.58	0.494	1	2
flag_sc	775,000	0.02	0.132	0	1

The full sample set consists of 775,000 samples

Table E.2: Summary statistics Monthly Performance File (svcgfile) Sample 2001-2016

Statistic	N	Mean	St. Dev.	Min	Max
svcg_cycle	35,822,926	201,037.8	391.585	200,102	201,612
current_upb	35,822,926	162,925.5	98,413.7	0.000	1,144,000
delq_sts	35,822,924	0.221	2.251	−1	124
loan_age	35,822,926	36.73	32.037	0	190
mths_remng	35,822,926	295.09	73.050	−31	603
repch_flag	515,723	0.006	0.079	0	1
flag_mod	35,822,926	0.0004	0.020	0	1
cd_zero_bal	515,508	1.086	0.463	1	4
dt_zero_bal	515,508	201,005.3	408.299	200,102	201,612
current_int_rt	35,822,926	5.397	1.124	0.000	20.000
non_int_brng_upb	35,822,926	157.65	3,605.50	0.000	332,500.0
dt_lst_pi	34,131	201,044.2	382.592	200,102	201,610
mi_recoveries	17,888	11,500.31	23,281.6	0	295,979
non_mi_recoveries	17,888	7,669.16	31,831.7	−48,920	511,239
expenses	17,888	−15,459.51	15,239.6	−158,583	244,886
legal_costs	17,888	−3,378.13	2,551.6	−70,304	0
maint_pres_costs	17,888	−4,902.75	7,403.17	−89,012	989
taxes_ins_costs	17,888	−6,361.62	9,430.71	−106,865	258,903
misc_costs	17,888	−660.044	3,803.12	−142,559	236,051
actual_loss	17,888	−72,185.60	62,593.1	−525,560	112,666
modcost	774,948	318.427	3,902.93	−15,427.	197,681.1

The full sample set consists of 35,822,926 samples

F. Results: Plots and Tables

Tables F.1-F.4 show tables that have been used to estimate to base our variables upon in the MNL model.

Table F.1: Results MNL Model 2001 with $\mathbb{P}$[On Schedule] ($k = 1$) vs. $\mathbb{P}$[Default], $\mathbb{P}$[Delinquent], $\mathbb{P}$[Part. Prepayment] and $\mathbb{P}$[Full Prepayment] ($\kappa = 2, \ldots, 5$) FICO, FirstHome, Mortgage Insurance, Number Of Units, Occupancy Status, CLTV, DTI, Original UPB, LTV, Interest Rate, Prepayment Penalty, Loan Term & Number of Borrowers.

	Dependent variable:			
	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
FICO Score	−0.012***	−0.013***	0.003***	0.002***
	(0.0001)	(0.00002)	(0.00005)	(0.00002)
First Home	−0.392***	−0.291***	0.215***	0.040***
	(0.00000)	(0.000)	(0.00000)	(0.00000)
Mortgage Insurance	0.007***	0.005***	0.003***	−0.002***
	(0.00000)	(0.00000)	(0.00004)	(0.00000)
Number Of Units	0.383***	−0.0003***	0.229***	−0.302***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Occupancy Status	−0.084***	−0.218***	−0.014***	−0.190***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
CLTV	−0.014***	−0.029***	−0.023***	0.004***
	(0.00001)	(0.00000)	(0.0001)	(0.00000)
DTI	0.004***	0.009***	−0.016***	−0.002***
	(0.00000)	(0.00000)	(0.00001)	(0.00000)
OrigUPB	−0.00000***	−0.00000***	−0.00001***	0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
LTV	0.038***	0.049***	0.021***	−0.005***
	(0.00001)	(0.00000)	(0.0001)	(0.00000)
InterestRate	0.715***	0.449***	−0.458***	0.463***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
PrepPenalty	0.827***	−0.017***	2.174***	0.594***
	(0.000)	(0.000)	(0.000)	(0.000)
LoanTerm	−0.045***	−0.057***	−0.032***	−0.036***
	(0.00005)	(0.00001)	(0.0001)	(0.00001)
Number Of Borrowers	−0.672***	−0.498***	−0.059***	0.086***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Constant	11.360***	22.300***	13.110***	5.021***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
AIC		1,519,075		
Note:	*Std. Errors in parentheses*		*p<0.1; **p<0.05; ***p<0.01	

Table F.2: Results MNL Model 2001 with $\mathbb{P}$[On Schedule] ($k = 1$) vs. $\mathbb{P}$[Default], $\mathbb{P}$[Delinquent], $\mathbb{P}$[Part. Prepayment] and $\mathbb{P}$[Full Prepayment] ($\kappa = 2, \ldots, 5$) FICO, FirstHome, Mortgage Insurance, Number Of Units, Occupancy Status, CLTV, DTI, LTV, Interest Rate, Prepayment Penalty, Loan Term & Number of borrowers.

	Dependent variable:			
	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
FICO	-0.012^{***}	-0.013^{***}	0.002^{***}	0.002^{***}
	(0.001)	(0.0001)	(0.00005)	(0.0001)
First Home	-0.375^{***}	-0.282^{***}	0.240^{***}	0.027
	(0.00005)	(0.017)	(0.007)	(0.018)
Mortgage Insurance	0.009^{**}	0.006^{***}	0.007^{***}	-0.006^{***}
	(0.004)	(0.001)	(0.0003)	(0.001)
Number Of Units	0.322^{***}	-0.055^{***}	0.126^{***}	-0.122^{***}
	(0.00002)	(0.002)	(0.011)	(0.001)
Occupancy Status	-0.057^{***}	-0.197^{***}	0.034^{***}	-0.245^{***}
	(0.0001)	(0.019)	(0.006)	(0.017)
CLTV	-0.015^{***}	-0.031^{***}	-0.030^{***}	0.009^{***}
	(0.003)	(0.003)	(0.001)	(0.002)
DTI	0.003	0.007^{***}	-0.019^{***}	0.001^{***}
	(0.003)	(0.001)	(0.0002)	(0.001)
LTV	0.037^{***}	0.051^{***}	0.023^{***}	-0.007^{***}
	(0.002)	(0.003)	(0.001)	(0.002)
Interest Rate	0.777^{***}	0.510^{***}	-0.261^{***}	0.301^{***}
	(0.0001)	(0.011)	(0.006)	(0.014)
Prepayment Penalty	0.799^{***}	-0.076^{***}	2.241^{***}	0.524^{***}
	(0.00000)	(0.00002)	(0.00001)	(0.00001)
Loan Term	-0.058^{***}	-0.058^{***}	-0.035^{***}	-0.033^{***}
	(0.002)	(0.0004)	(0.0002)	(0.0004)
Number Of Borrowers	-0.719^{***}	-0.545^{***}	-0.198^{***}	0.206^{***}
	(0.00004)	(0.012)	(0.005)	(0.012)
Constant	15.750^{***}	22.140^{***}	12.680^{***}	5.225^{***}
	(0.00000)	(0.00001)	(0.00001)	(0.00001)
AIC		1,519,075		
Note:	*Std. Errors in parentheses*		*p<0.1; **p<0.05; ***p<0.01	

Table F.3: Results MNL Model 2001-2005 with $\mathbb{P}$[On Schedule] ($k = 1$) vs. $\mathbb{P}$[Default], $\mathbb{P}$[Delinquent], $\mathbb{P}$[Part. Prepayment] and $\mathbb{P}$[Full Prepayment] ($\kappa = 2, \ldots, 5$) FICO, Mortgage Insurance, Original UPB, CLTV, DTI, LTV, Prepayment Penalty

	Dependent variable:			
	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$
FICO	-0.005^{***}	-0.012^{***}	0.003^{***}	0.0002^{***}
	(0.00003)	(0.00001)	(0.00000)	(0.00001)
Mortgage Insurance	0.004^{***}	0.007^{***}	0.004^{***}	0.003^{***}
	(0.00000)	(0.00000)	(0.00001)	(0.00000)
Original UPB	-0.00000^{***}	-0.00000^{***}	-0.00000^{***}	0.00000^{***}
	(0.00000)	(0.00000)	(0.000)	(0.00000)
LTV	0.048^{***}	0.024^{***}	0.005^{***}	0.001^{***}
	(0.00000)	(0.00000)	(0.00002)	(0.00000)
CLTV	-0.015^{***}	-0.009^{***}	-0.008^{***}	-0.001^{***}
	(0.00000)	(0.00000)	(0.00002)	(0.00000)
DTI	0.017^{***}	0.013^{***}	-0.014^{***}	-0.005^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Prepayment Penalty	1.955^{***}	1.012^{***}	0.273^{***}	0.391^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	-5.970^{***}	3.912^{***}	-1.744^{***}	-3.929^{***}
	(0.00000)	(0.000)	(0.00000)	(0.000)
AIC	18,674,167			
Note:				*p<0.1; **p<0.05; ***p<0.01

Table F.4: Results MNL Model 2001 with subsegmentation in κ=4 FICO & LTV

	Dependent variable:						
	$\kappa = 2$	$\kappa = 3$	$\kappa = 4.1$	$\kappa = 4.2$	$\kappa = 4.3$	$\kappa = 4.4$	$\kappa = 4.5$
FICO	0.0003	−0.0001	−0.0002	-0.002^{***}	0.001^{**}	-0.004^{***}	−0.0003
	(0.0002)	(0.0002)	(0.0002)	(0.0004)	(0.001)	(0.001)	(0.001)
LTV	0.001	−0.001	−0.003	0.0003	−0.007	-0.010^{*}	0.009
	(0.002)	(0.002)	(0.002)	(0.004)	(0.005)	(0.005)	(0.007)
Constant	2.645^{***}	3.033^{***}	0.699^{***}	-0.470^{***}	-2.342^{***}	1.331^{***}	-3.006^{***}
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00002)	(0.00002)

	Dependent variable:					
	$\kappa = 4.6$	$\kappa = 4.7$	$\kappa = 4.8$	$\kappa = 4.9$	$\kappa = 4.10$	$\kappa = 5$
FICO	0.001^{*}	0.0003	-0.003^{***}	0.0002	-0.001^{***}	0.009^{***}
	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0003)	(0.002)
LTV	−0.004	0.014^{*}	0.013^{*}	−0.002	−0.001	−0.015
	(0.006)	(0.007)	(0.007)	(0.005)	(0.002)	(0.016)
Constant	-3.161^{***}	-3.919^{***}	-1.598^{***}	-1.856^{***}	0.614^{***}	-9.817^{***}
	(0.00001)	(0.00002)	(0.00003)	(0.00001)	(0.00001)	(0.00002)
AIC:			113,695			

Note: *Std. Errors in parentheses* *p<0.1; **p<0.05; ***p<0.01

Figures F.1-F.5 show Markov transition probabilities over time. For each month a probability matrix is estimated. Each point is plotted over time, such that we can observe clearly in which time interval transition probabilities were smaller or bigger. Since the transition probabilities of $\pi_{i4,t}$ are already given in Figure 6.3 they are left out in each Figure.

Figure F.1: Markov Transition Probabilities $\pi_{1j,t}$ for $j = 1, 2, 3, 5$ and $t = 9, \ldots, 192$

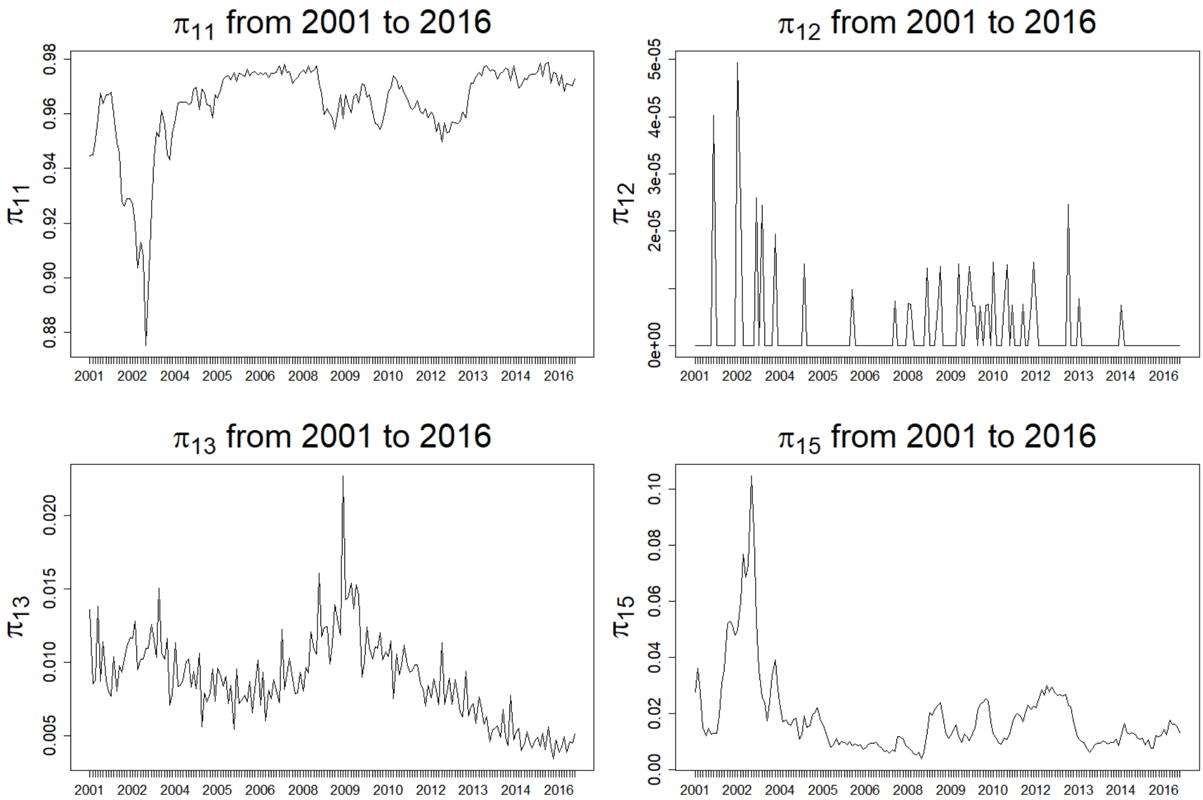


Figure F.2: Markov Transition Probabilities $\pi_{2j,t}$ for $j = 1, 2, 3, 5$ and $t = 9, \ldots, 192$

Figure F.3: Markov Transition Probabilities $\pi_{3j,t}$ for $j = 1, 2, 3, 5$ and $t = 9, \ldots, 192$

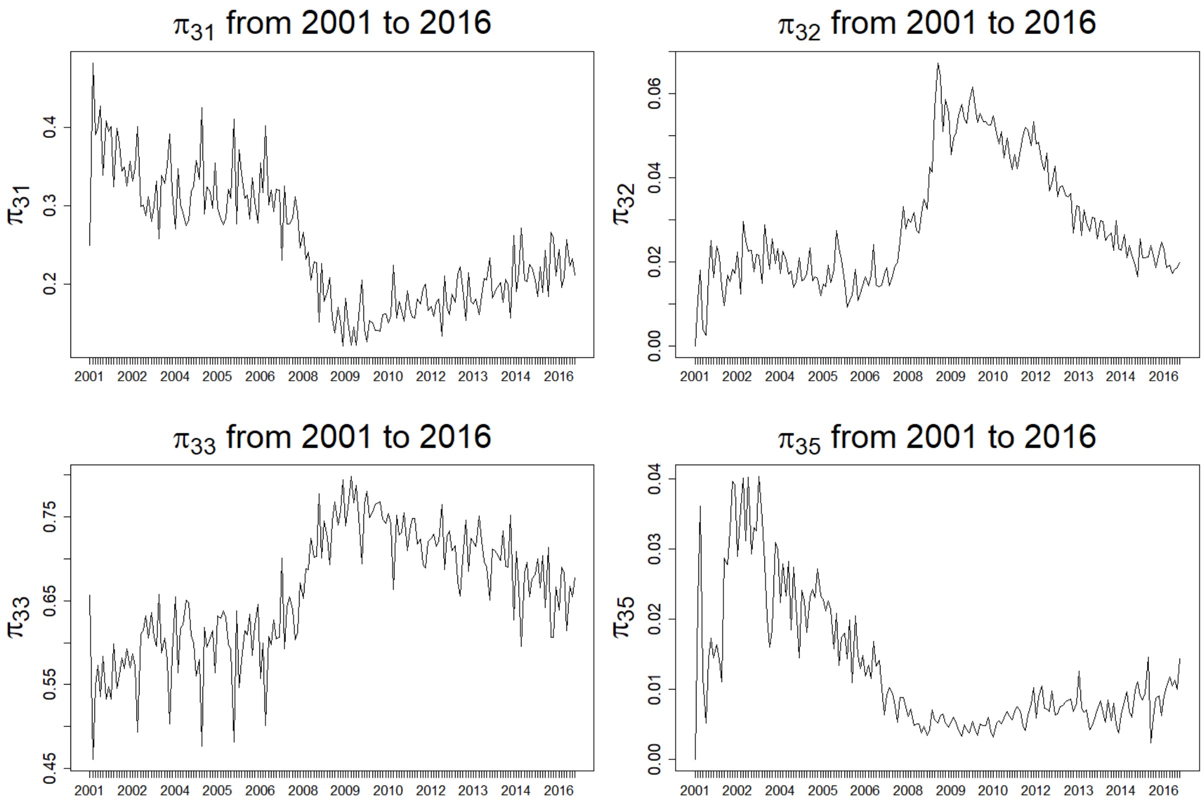


Figure F.4: Markov Transition Probabilities $\pi_{4j,t}$ for $j = 1, 2, 3, 5$ and $t = 9, \ldots, 192$

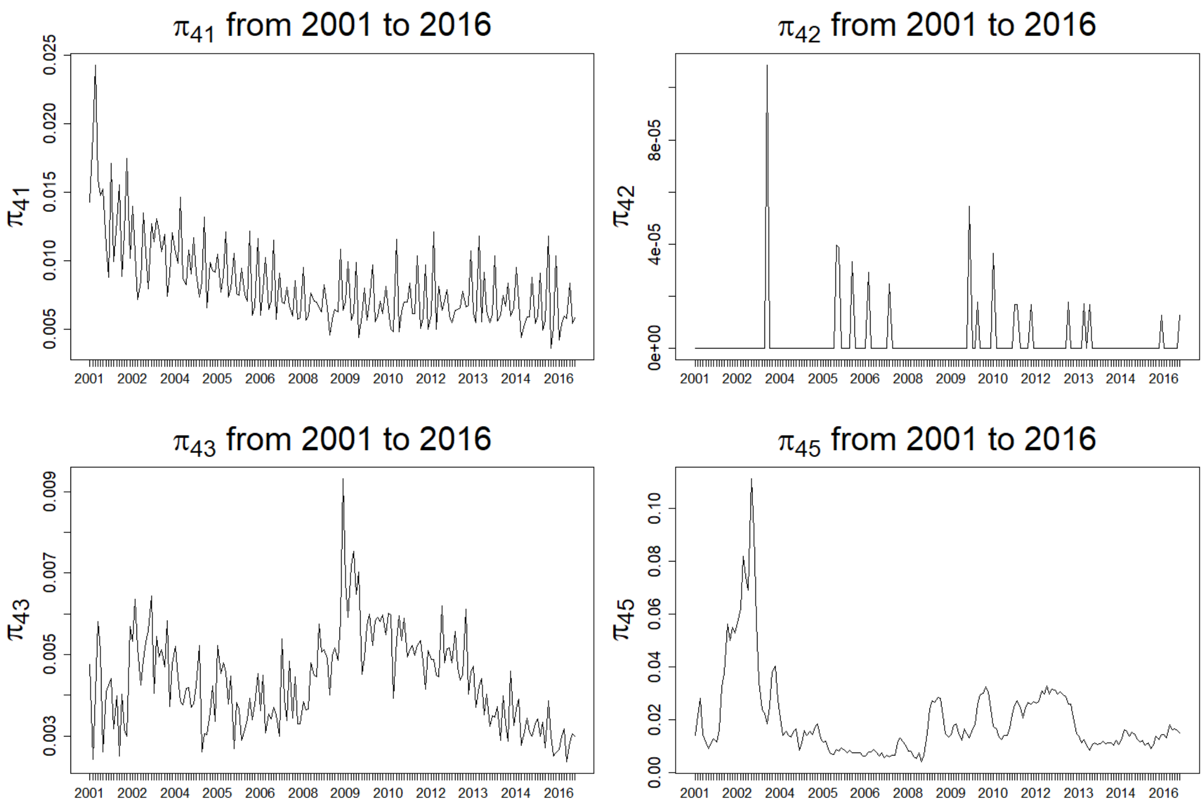


Figure F.5: Markov Transition Probabilities $\pi_{5j,t}$ for $j = 1, 2, 3, 5$ and $t = 9, \ldots, 192$