Erasmus University Rotterdam

Master Thesis

Accounting for heterogeneity in aggregated mortgage pools with the aim of forecasting prepayment / default risk

Author:
Gert de Rijke (351544)

Supervisor:
Bart Keijsers

August 19, 2016

Abstract: I investigate the use of segmentation of aggregated mortgage pool data in a Markovian setting. I aim to increase out-of-sample forecast accuracy of mortgage performance. To this end I implement a Bayesian multinomial logit model with an adaptive Metropolis-Hastings algorithm to sample the posterior distribution. I devise a decision rule to map the multinomial probabilities to mortgage performance forecasts. I find that segmentation based on mortgage characteristics not directly defined in the data set, but that can be properly formulated, increases model fit and forecast accuracy for non-performing loans. I also discover that performing loans demonstrate a remarkable level of homogeneousness.

Contents

1 Introduction

The mortgage market is a size-able part of any modern economy. Mortgage providers can consist of a vast array of different entities. From government backed organisations, large banks or insurance companies, to smaller local mortgage providers. To determine the value of a mortgage some factors play key roles. Consider a mortgage to be a financial product that gives the mortgage provider the right to receive cash flows in the future consisting of principle repayment and interest payments. It then follows that the value of such a product can be related to the net present value of the future cash flows. Ideally the cash flows would be deterministic and we would only have to worry about determining the discount rate. Sadly, for mortgages these future cash flows are not certain. Broadly, two major events can take place which greatly determine the cash flows. Namely, a prepayment event and a default event.

A prepayment event is when a mortgage borrower repays the remaining principle of the mortgage before the end of the legal maturity. A default event is when the mortgage borrower will no longer repay principle or interest on the mortgage, effectively relinquishing his/her home. When a mortgage borrower prepays their mortgage it effectively terminates the interest payments, as well as generate a large cash flow of principle repayment. Likewise, a default event also ends the interest payments (of course a loss of principle repayment can also be incurred).

The main goal of my research is to forecast prepayment and default rates in a mortgage pool. As prepayment is closely related to the general performance of a mortgage, it makes sense to model the other performance events of a mortgage simultaneously. A performing mortgage is a mortgage the is up-to-date with its monthly payments and can be said to be CURRENT. A non-performing mortgage is one that is behind on its contractual payments, I consider two degrees of non-performance; DELINQUENT 30-89 DAYS and DELINQUENT 89+ DAYS, as is further discussed in Section 4.1. Mortgages can end their life as either having DEFAULTED or PREPAID. Modeling the evolution of a mortgage through these 5 states it becomes clear that, as the probability of one event occurring increases, the probability of other events occurring are also impacted. One can think of it as follows: The chance that a mortgage borrower will prepay their mortgage (be it refinancing or for another reason) declines as the probability of a default increases. These two events compete with each other. As the probability of one changes so will the probability of another.

To incorporate this interconnectedness of a mortgage prepaying and defaulting I start off with a Markovian structure wherein a mortgage can find itself in one of five defined states at each time period. This is further discussed in Section 4.1. The transition probabilities between the states of the Markov chain are modeled by a multinomial logit model. This allows for observable factors that effect the transition probabilities between the states to enter the model.

Prepayment and default are not deterministic events but probabilistic. Therefore many mortgage providers use the information in their mortgage portfolios to determine driving factors behind these events. A mortgage provider with a large and diverse portfolio has ample information to base their models on. For a smaller mortgage lender this quickly becomes problematic. As for this provider the own mortgage portfolio might not be large enough to produce accurate estimates for the parameters in their models. It is then not uncommon to make use of aggregated data on mortgage performances to construct a model. These aggregated data-sets can be made available by (semi) governmental institutions or privately run data vendors. The down side is that this does not account for the fact that the aggregated information might not be representative for the held mortgage portfolio. These unique properties could arise from the geographical location of the mortgage provider, or the unique underwriting mandate it implements. We can take this line of thought a step further and assume that aggregated mortgage data might not be homogeneous. In the sense that within

the aggregated mortgage pools there exist subgroups of homogeneous mortgage loans that share the same characteristics.

A simple example, consider a mortgage that was provided pre-financial crisis to a borrower with a bad credit history. This borrower could find it difficult to refinance (prepay) their mortgage at a later point in time when underwriting practices have become more strict. Where a decline or rise in interest rates normally effect the decision for a borrower to refinance their loan, this individual would not react in the same manner to movements in the interest rate. The fact that this borrower was subject to less stringent underwriting laws at origination is not captured in loan characteristics, but we can identify this sub-group by looking at origination year or possible geographical location.

The idea of my research is not to treat the entire data-set as being influenced in the same way by the explanatory variables. I propose to segment the data-set into smaller groups for which the argument can be made that the loans in each group more closely resemble each other than the entire available pool of mortgages. For example, consider the geographical area from which each loan originates. Due to differences in laws and lender practices, as well as differences in local economy and housing markets, it makes intuitive sense that loans from different geographical areas can relate differently to the explanatory variables. This segmentation of the total pool of mortgages into more homogeneous groups allows to capture the unobserved difference between mortgages. Unobserved in the sense that its effects are not captured by the available data. Section 4.3 goes into further detail on the proposed segmentation of the mortgages.

A down side of segmenting the data is that some of the resulting groups can end up having only a handfull of observed mortgages. This makes it difficult to estimate the parameters in the multinomial logit model. This is especially true for transitions that rarely occur. For example, the event of a performing mortgage suddenly going into the DEFAULT state is not something that occurs frequently. Discovering what drives such a transition requires enough observations to adequately estimate model parameters. Few observations could result in large parameter uncertainty, which in turn cause inaccurate forecasts. Misspecifying the relationship between risk drivers and the events they influence could result in under estimating financial consequences. In summary, allowing for a more granular modeling of mortgage performance is desired, but not when the increase in uncertainty in the to be estimated parameters can't be offset.

For this I make use of a Bayesian multinomial logit model. The idea is that the parameters estimated in the Bayesian multinomial logit model have a prior distribution centered on the parameter estimates of a multinomial logit model of the aggregated data pool. If a segment has few observations, the prior can be specified to be more informative and pull information from the larger pool of mortgages. If a segment has enough mortgages then the prior distribution will be set to be less informative. This setup allows to take advantage of modeling the non-homogeneity of mortgages, whilst overcoming the problem of uncertainty due to low number of observations. This empirical Bayes approach allows for the parameters of sub-segments of the mortgage pool to differ, but I do not expect these changes to be far off from the aggregated values. Section 4.5 goes into further detail of the Bayesian multinomial logit model. To estimate the posterior density of the Bayesian approach I implement an adaptive Metropolis-Hastings algorithm.

One difficulties with working with multinomial probabilities, as is the case with Markov chain transition probability matrices, is a way to evaluate the resulting forecasts, especially when some events have a relatively low probability of happening[1]. To this end I propose a novel way to map multinomial probability vectors to a single state transition. I do this whilst maintaining the aggregate properties of all estimated transition probabilities. This allows me to compare my forecasts to benchmark models. This decision rule is discussed

[1]Such as events whose probability is never forecast to be the highest probability event.

in Section 5.

A practical result of this framework comes from the experience of many (smaller) mortgage providers. Smaller mortgage providers often only observe a relatively small number of mortgages (their own) in real time. Estimating credit and prepayment risks can then be difficult. It is not uncommon for these smaller mortgage provider to use (historical) aggregate market data to determine these risks. A problem is that the unique characteristics of the mortgage providers own portfolio is not taken into account. It is not uncommon for different mortgage provider to maintain different underwriting standards, making aggregate date not as representative as might be desired. This is a problem as regulatory agencies prefer financial institutions to asses the risk based as much as possible on their unique set of assets.

Additionally a more general result of this framework is that it allows to examine the, assumed, heterogeneity of the aggregated mortgage pool. By segmenting into smaller homogeneous subsets the hope is to improve prepayment and default risk forecasts. The above leads to the following research question:

Does accounting for the heterogeneity in aggregated mortgage pools increase accuracy of forecasting mortgage prepayment/ default?

I focus mainly on heterogeneity in the geographical origination of mortgages in the main narrative of this thesis. In addition to this I also consider segmentation by vintage and a data driven segmentation by K-means algorithm. I start by giving an overview of existing literature in Section 2. Section 3 gives more detail of the used data-set and macro economic series, as well as provide some summary statistics. This is followed by Section 4 where an in depth explanation of the methodology is provided. Section 4.2 provides a description of the multinomial logit model, this is expanded upon with segmentation in Section 4.3. Section 4.4 outlines the procedure to generate the K-mean segmented data. I then continue with the specification of the Bayesian multinomial logit model in Section 4.5 and the sampling algorithm in Section 4.6. In Section 5 comparison measures are discussed before looking at the results in Section 6.

2 Literature

The modelling of prepayment behaviour of mortgages has acquired a substantial amount of interest over the past few decades. Especially for the American housing markets where it has become the standard for mortgage lenders to repackage their portfolios into mortgage backed securities (MBS). The US residential mortgage backed security market totaled $ 7.5 trillion at the end of 2013 (Campbell *et al.* (2014)). In most cases this concerns agency mortgage backed securities, that are guaranteed and/or repackaged by government backed enterprises such as the Federal National Mortgage Association (Fannie Mae), Federal Home Loan Mortgage Company (Freddie Mac), or government agencies such as the Government National Mortgage Association (Ginnie Mae).

The close relationship of the price of a MBS and prepayment rates becomes clear when considering that the mortgage borrower holds a call on the remaining future cash flows (principle and interest payments) with a strike at the current remaining principle. Alternatively default can be seen as a put option on the underlying home with a strike equal to the present value of the future cash flows. Kau & Keenen (1995) give a broader overview of this option theoretic approach. Under optimal conditions and rational behaviour a MBS could thus be priced as a callable bond. The holder of the mortgage would exercise the call when interest rates fell increasing the present value of the future cash flows above the nominal value of the mortgage.

Likewise, exercising the put when the price of the underlying home drop below the present value of the future cash flows. However, as Dunn & McConnell (1981) show, homeowners do not follow this rational behaviour and often times prepay when it is not optimal to do so.

This empirically observed irrational behaviour has motivated the distinction between interest rate related and non-interest rate related prepayment behaviour. Where interest rate related prepayment occurs when home owners refinance their mortgage to take advantage of lower interest rates. Non-interest rate related prepayment occurs when a sub optimal choice is made to prepay due to other factors, such as unemployment, job change, divorce, etc.

Downing *et al.* (2005) argues that broadly two approaches have been used to model prepayment behaviour. Reduced-form and structural models. With structural models it is assumed that prepayment and/or default behaviour is the rational behaviour of a mortgage borrower. These choices being driven by variables like interest rate and housing prices. Dunn & McConnell (1981) were the first to formulate such a model. They implemented the interest rate model of Cox *et al.* (1985) to price the implied options. Additionally they include irrational behaviour by Poisson process for suboptimal prepayment. They were the first to explicitly model the irrational behaviour in this option theoretic framework. Other structural models followed the work of Dunn & McConnell (1981). Note-able of which is the work of Stanton (1995). Stanton (1995) extends the option theoretic approach by including transaction costs. These transactions costs have a broader interpretation than just monetary costs of refinancing ones mortgage. These transaction costs can also be interpreted to include the perceived costs of prepayment by the homeowner. This helps explain the irrational behaviour of not prepaying when it is optimal to do so. Additionally, Stanton (1995) assumes that prepayment decisions occur at discrete times, as opposed to continuous. These additions help model the empirically observed burnout effect in mortgage pools, where prepayment rates decrease the longer it has been optimal to do so. The interpretation of this burnout effect in this framework is that homeowners who fail to prepay when it has been advantageous to do so apparently perceive a higher transaction costs (due to exogenous factors) that prohibit them from prepaying.

Within the reduced-form models the aim is to model prepayment as a function of possible explanatory variables. Historical information is used to estimate these models. Schwartz & Torous (1989) are one of the first to present a comprehensive model where prepayment is modeled as a function of variables such as seasonality, long term refinancing rates, and mortgage age. The relation between the explanatory variables and prepayment is modeled with a proportional hazards model estimated via maximum likelihood.

Further expanding on the seminal work of Schwartz & Torous (1989), Smith, Sanchez, and Lawrence (1996) take the reduced-form approach to prepayment and implement it in a Markovian framework. The idea is that the observed distinction in the behaviour of mortgage borrowers with respect to prepayment or default can, in part, be explained by the payment behaviour. A mortgage is modeled as being in a unique state indicating if the homeowner is up-to-date with his/her payments. Smith, Sanchez, and Lawrence (1996) define the states current, delinquent, payed-off, and default. The influence of considered explanatory variables on the prepayment and default rate is then modeled separately for each considered state with a multinomial logit model. Intuitively, the advantage of this framework is that it allows to model the observed fact that a homeowner that has a higher chance to default (i.e. is delinquent in payments) has a lower chance to suddenly prepay the mortgage. Grimshaw & Alexander (2011) utilize the same Markovian framework but impose a prior distribution on the state transition probabilities to allow exogenous information to enter the model. Both papers find that the Markovian structure improves the modeling of prepayment rates over that of a stateless model such as that of Schwartz & Torous (1989). On a side note, the modeling of prepayment and

default rates simultaneously, as is done with the Markovian structure, is intuitive from the perspective of a MBS investor. Zipkin (1993) points out that from the view of an investor in government guaranteed mortgage backed securities a default on a mortgage has broadly the same implications as a prepayment. Namely, a discontinuation of future interest rate payments, as the nominal value of the mortgage is guaranteed.

Where it has been standard to model the relationship between prepayment rates and considered explanatory variables via discrete choice models (e.g. logit/probit) recent research has focused on the use of a Bayesian framework. Popova *et al.* (2008) propose the use of a Bayesian Mixture model in the proportional hazard framework of Schwartz & Torous (1989) to account for differing amounts of prepayments in GNMA mortgage pools. This allows to differ between a constant rate of sub-optimal prepayment and larger, sudden, prepayment rates caused by refinancing. Popova *et al.* (2008) find that a two mixture model (i.e. bi-modal distribution) improves prepayment forecasting over the uni-modal variant. Grimshaw & Alexander (2011) also utilize a Bayesian framework. They apply a Dirichlet prior to the transition probabilities in a Markovian structure when modeling small groups of mortgages. The prior allows possibly relevant information from aggregate data to enter the model for smaller mortgage subsets. These smaller mortgage portfolios are often times found with smaller mortgage providers. This setup allows the use of possibly unique but limited information in the smaller mortgage portfolio to estimate model parameters. This information is augmented by information from aggregate, although possibly less relevant, data to improve estimation accuracy. Lastly, Sung, Soyer, & Nhan (2007) implement a Bayesian multinomial logit model to model non-homogeneous Markov chains. Although their paper is not directly related to mortgages, their theoretical framework can be applied to the Markovian structure of Smith, Sanchez, and Lawrence (1996). Where Grimshaw & Alexander (2011) specify a prior on the transition probabilities of the Markov chain directly, the approach of Sung, Soyer, & Nhan (2007) allows to specify a prior on the parameters of the multinomial logit model that describe the relationship between the transition probabilities and the explanatory variables. The parameters are then estimated with the use of a Gibbs sampler.

3 Data Description

The data I use in my research is made available by the Federal Home Loan Mortgage Corporation, also known as Freddie Mac. Freddie Mac is a government sponsored enterprise in the United States of America. Together with the Federal National Mortgage Association, Fannie Mae, it was established to increase liquidity in the mortgage market. Both Freddie Mac and Fannie Mae purchase mortgages on the secondary market and pool these mortgages together. These pools are then sold as (residential) mortgage backed securities, often denoted as RMBS, where the cash-flows originating from a pool of mortgages is divided over the buyers of the securities. Freddie Mac and Fannie Mae often guarantee the timely repayment of principle and interest on the mortgages in these pools, as such taking on the credit risk in these products. For this they would charge a premium as compensation. The purchasing of mortgages from mortgage providers and essentially pooling the credit risk allows for matching of risk averse (institutional) investors to the originators of mortgages. This increases liquidity affordability of home ownership in the United States.

Although Freddie Mac and Fannie Mae are government sponsored enterprises, the federal government does not provide any backing of the guarantees that Freddie Mac and Fannie Mae provide. This in contrast to the Government National Mortgage Association, Ginnie Mae, which is backed by the federal government. Due to the nature of the business of Freddie Mac and Fannie Mae these enterprises accrue a lot of data on mortgage performance. In an effort to increase transparency and provide other institutions with data to

increase there credit models, Fannie Mae and Freddie Mac make there mortgage performance data available.

The Freddie Mac data-set consists of 22 million single family 30 year fixed rate mortgages that have been originated between 1999 and 2015, covering all 50 states. It is maintained and updated by Freddie Mac on a regular basis. For ease of use Freddie Mac has sampled $50,000$ mortgages per origination year from the mortgages it has purchased or has guaranteed.

I concentrate my research on the California mortgage market. The main advantage of this is that California is the largest geographical state by population, resulting in a large number of observations. Additionally it allows me to more easily compare results to those published by Smith, Sanchez, and Lawrence (1996) whom also studied the Californian mortgage market. Lastly restricting the research to one geographical state mitigates possible differences in legislature in underwriting practices between states.

A big advantage of the Freddie Mac single family data-set over, for example, European data-sets (from the European Data Warehouse) or data-sets originating directly from mortgage providers, is the granularity and completeness of the included information. Information acquired from mortgage providers is generally hard to come by and of significantly smaller size. European data tends to be incomplete (missing values) and not as detailed as the Freddie Mac data-set.

Although mortgages are available all over the world, the characteristics often change dependent on the region. As a result, analysis of American mortgages might not be directly applicable to the European (or Dutch) mortgage markets. In general the American mortgage market contains some differences that make direct application of results gotten in this data-set tot he European (Dutch) market inadvisable. Most notable of these differences is the, under most circumstances, penalty free prepayment of the mortgage. Which is currently not possible in the Netherlands. This means that a large factor of financial motivated behaviour, i.e. refinancing when optimal, is not applicable to the Dutch market. Although the driving factors behind PREPAYMENT and DEFAULT will always differ per country and region[2], the way of modeling the drivers effecting these risk factors can be applied in different circumstances. In other words, the covariates and how these relate to the mortgage performance will change between countries. The estimation of these effects, as discussed in this research, can be generalized.

I additionally make the choice to concentrate on estimating the considered models over 2010 data and forecasting performance of the mortgages in 2011. This decision was made as the data-set contains diverse observations for different types of mortgages during 2010. There are still a sufficient number of mortgages from pre-2007, negative equity, young, jumbo, and other characteristics. Earlier years the data-set is still growing, as it was started in 1999. Later years contain fewer observations of mortgages originated at the turn of the millennium as a lot of these have either prepaid or defaulted. The set-up of my research allows for the modeling of the unique characteristics of mortgage from different origination years, see Section 4.3, as such this approach remains generalizable.

The mortgage loan information available in the data-set falls into two categories, origination and monthly performance data:

Origination data: This consists of loan level data at the time of origination of the mortgage. This encompasses borrower information as well as property and mortgage attributes. Examples are CREDIT SCORE, PROPERTY TYPE, and POSTAL CODE. The POSTAL CODE is made available as only the first 3 digits for privacy reasons. This is however enough to match each mortgage with the county it is in, which in turn allows the addition of loan level macro economic variables as explained below.

[2]In fact, part of the subject of this research is that not only do these characteristics differ on a macro geographical level, but also on a more local geographical level.

Monthly performance data: This encompasses the loan level variables that are tracked on a monthly basis by Freddie Mac. Examples are CURRENT UNPAID BALANCE, LOAN DELINQUENCY STATUS, LOAN AGE, etc. These series are available monthly up to and including the month of the loans termination event.

The Tables 20 and 21 in the Appendix contain an exhaustive list and description of the series in the Freddie Mac single family 30-year fixed rate data-set. The termination of a mortgage is noted in the data-set as the occurrence of one of four possible events:

1. Prepaid or Matured (Voluntary Payoff)
2. Foreclosure Alternative Group (Short sale, Third Party Sale, Charge Off or Note Sale)
3. Repurchase prior to property disposition
4. REO (Real estate owned) Disposition

Of the four stated possible events the first is considered a prepayment event, where as the latter three are considered a mortgage default scenario. Events 2 and 3 are a default that result in the property being transferred to a third party. Event 4 is a default scenario where Freddie Mac took direct ownership of the property. Note that although the first event can also imply that the mortgage matured, given that the data-set consists of 30-year mortgages which originated since 1999 this can realistically never be the case. Appendix A.3 includes a complete list of the available series in the Freddie Mac data-set. In addition to the Freddie Mac data-set I used the macro-economic series that are publicly available from the Bureau of Labor Statistics and the Federal Reserve of St. Louis:

1. County level housing price index; Bureau of labor statistics
2. County level unemployment rates; Bureau of labor statistics
3. Postal code population numbers; Bureau of labor statistics
4. Market average 30-year fixed mortgage spot rates; Federal Reserve of St. Louis

These series, in combination with the 3 digit postal code per mortgage in the Freddie Mac data-set, allowed for the matching of macro-economic series to each mortgage loan. Section 3.2 goes into more detail which series, and their transformations, are considered as explanatory variables. With respect to missing values the data series obtained from the Bureau of labor statistics and the Federal Reserve of St. Louis are of high quality and did not contain any. For the Freddie Mac data-set if a mortgage had a missing value for any of the required series the mortgage was excluded from the analysis. This resulted in 144 mortgages being removed from the data-set[3].

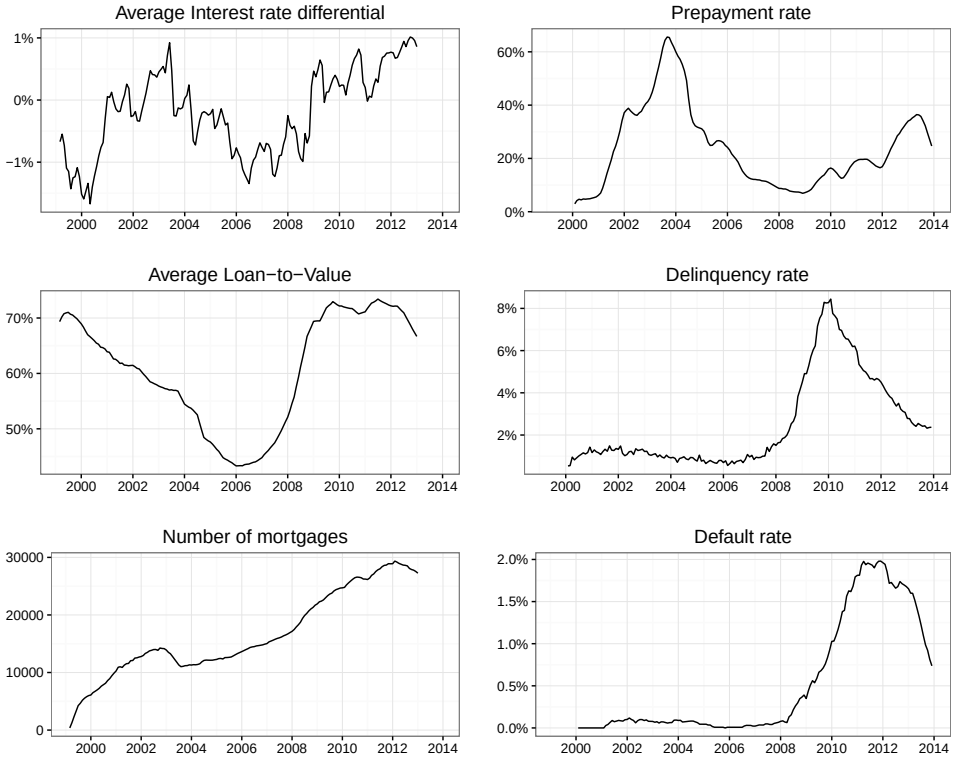
The data will be initially split into mortgages that are up-to-date with their monthly payments (CURRENT), mortgages that are between 1 and 3 months behind on their payments (DELINQUENT 30-89 DAYS), and mortgages that are more than 3 month behind on payments (DELINQUENT 89+ DAYS). Also, mortgages can be denoted as having transitioned to DEFAULT or PREPAID in some period. Mortgages that are up-to-date with their payments can also be called performing mortgages, in contrast to mortgages that are behind on payments or have defaulted, which can be called non-performing mortgages. Section 4.1 goes into more detail on the implied underlying Markovian structure.

[3]The missing value check was preformed after the explanatory variables for each model had been determined, as explained in Section 4.2, as to not to exclude mortgages from the data set that only lacked values in data series that would not have been required.

3.1 Summary statistics data-set

Figure 3.1 below shows the performance of Californian mortgages in the data-set. The stated prepayment and default rates are calculated over a moving window over the previous 12 months. As such, the stated percentages are the proportion of mortgages that have prepaid or defaulted in the past 12 months, conditional on them existing at the start of those 12 months. Additionally the average interest rate differential, Loan-to-Value (LTV) ratio, and the number of mortgages in the sample are shown.

Figure 3.1: Californian mortgage performance 1999 to 2013



Prepayment, delinquency, and default rates are calculated as the proportion of mortgage loans that have prepaid or gone into default in the previous moving year. Interest rate differential is the difference between that mortgage rate on a loan and the average market mortgage rate at the stated time.

It can quickly be noted that prepayment rates have historically varied greatly. From a prepayment rate of below 10% in 2009 to more than half of the mortgages prepaying their nominal value in 2003. A large component in prepayments has been the possibility for a borrower to refinance their loan at a lower rate. This effect can be seen when looking at the average interest rate differential, indicating the difference between the rate on a mortgage and the average market mortgage rate. In times where the average difference between the rate on existing mortgages and the spot mortgage market rate are increasing (such that refinancing is becoming financially attractive) one can also see a sharp increase in prepayment rates. This observed relationship between interest rate differential and prepayment rates is in-line with existing research. Long term interest rates, for their influence on offered mortgage rates, are therefore often seen as the most important driver of prepayment rates (Dunn & McConnell (1981), Schwartz & Torous (1989) to name two).

It can be seen that the default rate of the mortgages in the data-set starts stable until mid 2008. It

should be noted that the mortgages in the data-set consist primarily of prime (and conforming) mortgages and not the, now well known, sub-prime category. At a closer look it can be seen that default rates are preceded by a sharp increase in loan-to-value ratios, caused by a significant decrease in housing prices[4]. This results in increased possibility that the property value underlying the mortgages is less valuable than the outstanding notional of the mortgage loan (also known as negative equity). A borrower would have reduced options if they came to be unable to afford the monthly mortgage payments. Where refinancing or selling the property could have offered a solution in the case of positive equity, with negative equity a borrower can be incentivized to default on their mortgage[5].

3.2 Explanatory variables

From the Freddie Mac data-set and the mentioned macro-economic series, explanatory variables are taken or constructed to model the development of the performance of the mortgages. The multinomial logit model I implement is further explained in Section 4.2. Here I outline the considered covariates that enter the model and, if a transformation was applied, how these have been constructed.

Table 1: Explanatory variables

Variable	O/M	Type	Range
Provided			
Credit Score	O	numerical	[456, 840]
Original LTV	O	numerical	[6, 100]
Original DTI	O	numerical	[0, 65]
Occupancy Status	O	categorical	{I, O, S}
Number of borrowers	O	categorical	{1,2}
Channel	O	categorical	{B,C,R,T}
First time homebuyer	O	categorical	{True, False}
Loan purpose	O	categorical	{C, N, P}
Constructed			
Log(loan age)	M	numerical	[0, 4.875]
Every delinquent	M	categorical	{True, False}
Under water	M	categorical	{True, False}
Interest rate differential	M	numerical	[-3.53, 3.96]
Time since profitable	M	numerical	[0, 116]
Estimated LTV	M	numerical	[0.39, 194.33]
County Unemployment Rate	M	numerical	[3.8, 13.8]
Jumbo	O	categorical	{True, False}

A list of considered covariates to model the performance development of mortgages. O/M indicates if it is a (M)onthly observations or only at (O)rigination. Range indicates the set of possible values the covariate can take.

The explanatory variables that have been considered when fitting the multinomial Logit model can be found in the Table 1. Variables that have had no transformation applied to their value are considered as "Provided". These are are covariates from the Freddie Mac data-set and are mortgage loan characteristics. "Constructed" variables are covariates which have been created by combining attributes of individual mortgage loans with macro economic data or applying some transformation to the provided series. Variables

[4]Another possible explanation for rising LTV ratios is an increase of new mortgages in the data-set. This is because new loans tend to have higher LTV ratios. This is however not the case, as there is no significant increase in inflow, let alone enough to have the average LTV ratio jump by more than 30%.

[5]Although this does have other repercussions such as a drop in credit rating and possibility of legal action

denoted with M have monthly observations, where as variables denoted with O are observed only at the origination of the mortgage loan. Following is a short description of each considered explanatory variable.

Credit score A score provided by a third party credit agency at the time of the mortgage origination indicating the borrower's creditworthiness. Figure 3.2 shows the credit score in different transition scenarios. I can be seen that mortgages that transition to PREPAID tend to have higher credit scores than other mortgages. This is possibly a result of the fact that a higher credit score increases the possibility of being accepted for a refinance loan.

Original debt-to-income (DTI) Is the ratio of total debt payments over the stated income at the time of origination of the mortgage. Increasing DTI ratios in originated loans in the years prior to 2007 has been seen as one of the drivers that resulted in the sub-prime crisis (see for example Mian & Sufi (2009). DTI can also be seen as a measure of a mortgage borrowers ability to fulfill their obligations. Where DTI is higher, the probability of payment problems increases as the borrower is less capable of handling economic setbacks. In Figure 3.2 it can be seen that mortgages that transition to one of the non-performing states tend to have a slightly higher DTI than mortgages that stay CURRENT or PREPAY.

Occupancy status Possible values are *Owner occupied (O)*, *Investment property(I)*, and *Second home (S)*. The varying occupancy categories can have different financial situations leading to different priorities in their debt payments. If a property is a *second home* it could indicate that the owner is financially better off than the average. As such better able to keep up with monthly payments or refinance the mortgage should it be profitable to do so.

Number of borrowers Differentiates between mortgages with 1 borrower (indicated by 1) and those with more than 1 (indicated by 2). More borrowers signing on the loan offers more certainty for the lender that the mortgage will be repaid.

Channel Indicates the channel via which the loan was originated. *Broker (B)*: An entity that originates mortgages by matching borrowers with lenders, and receiving a commission for this service. Brokers never hold the mortgages on their books. *Correspondent (C)*: An entity that originates mortgages and holds these temporarily prior to selling to a third party. *Retail (R)*: Entity that originates and holds the mortgage themselves. *Not specified*: If unknown which channel the mortgage originated via. Different origination channels can have different underwriting criteria.

First time home buyer True or false to indicate if the borrower has ever owned a property before.

Loan purpose *Purchase (P)*: The mortgage was provided to purchase a home. *No cash-out refinance (N)*: The mortgage was a refinancing where no equity was removed. *Cash-out refinance (C)*: The mortgage was a refinancing where equity was removed from the home.

Log(loan age + 1) The age of a loan (in months) is seen as an important indicator for prepayment risk, especially early in the life of a mortgage. It is observed that prepayment rates early in the life of a mortgage are generally lower due to the fact that home buyers are unlikely to buy a home if they expect to become unemployed, change job, or undergo other influential factors that would cause them to prepay or default on a mortgage (see Smith, Sanchez, and Lawrence (1996) and Hayre & Young (2004)). Taking the logarithm allows the effect of this variable to decrease for longer vintages but still play an important role in the short term.

Ever delinquent This variable indicates if the mortgage, over its life up to and including t, has ever been behind on payments. Naturally this can only not be the case for mortgages that are CURRENT on payments.

Underwater Indicates if the mortgage notional is more than the property value at time t, also known as negative equity.

Interest rate differential The general consensus in literature is that interest rate differential is the single biggest explanatory variable for (optimal) prepayment in mortgages. From the contingency claims models originating with Dunn & McConnell (1981) through the the proportional hazard model of Schwartz & Torous (1989) and more recent Bayesian Mixture model of Popova *et al.* (2008) the difference between the interest rate on a mortgage and the market mortgage rate has proven of significant explanatory value. The interest rate differential at time t for mortgage m is defined as the mortgage rate $r_{m,t}$ minus the average market rate, as published by the Federal Reserve of St. Louis, at the time; $\bar{r}_{market,t}$.

$$\Delta r_{m,t} = r_{m,t} - \bar{r}_{market,t} \tag{3.1}$$

Time since profitable It has been noted in previous research (Popova *et al.* (2008) and Schwartz & Torous (1989)) that the more time has passed since it was economically efficient for a borrower to prepay their mortgage, the less likely it becomes that they will. A possible explanation of this effect is that a borrower might not be able to refinance their mortgage due to extraneous reasons (e.g. cannot meet underwriting requirements, bad credit score, etc.). Time since profitable for some mortgage m at time t is then $C_{m,t}$:

$$C_{m,t} = \begin{cases} C_{m,t-1} + 1 & \text{if } \Delta r_{m,t} > 0,5 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

Where a cost of 0.5% for refinancing is implied. Stanton (1995) investigate in more detail the financial and non-financial perceived transaction costs. Where they model transactions costs as stochastic, they find it distributed between 30% to 50% of the nominal value. Assuming a remaining time to maturity of 25 years, a constant transaction costs of 0.5% is a justifiable simplification.

Estimated Loan-to-Value (LTV) Is the ratio of outstanding unpaid balance to the estimated value of the property. First the original property value was calculated using the unpaid balance and the LTV ratio at origination. Each loan was matched with a county based on the 3-digit postal code available in the data set. Then using a county level housing price index the monthly property value can be estimated. The estimated LTV for some mortgage m at time t is then the unpaid balance at t divided by the estimated property value:

$$\hat{LTV}_{m,t} = UPB_{m,t} \left(\frac{hpi_{m,t}}{hpi_{m,o}} \frac{UPB_{m,0}}{LTV_{m,0}} \right)^{-1} \tag{3.3}$$

Note that a high $\hat{LTV}_{m,t}$ can be caused by declining housing prices (i.e. $hpi_{m,t}$ declines) but also by a high $LTV_{m,0}$ at origination. As LTV increases a borrower might become less willing to pay as the payments go toward the redemption of a loan where the underlying collateral is valued less than the outstanding notional. Indeed if we look at Figure 3.2 we notice that the majority of mortgages that

ended the year as non-performing have LTV ratios $> 80\%$ (the red dots). Especially when considering that the average LTV ratio for the period is around 70%, meaning that the higher proportion of non-performing mortgages is not due to a generally higher proportion of $> 80\%$ LTV loans.

County Unemployment rate The level of unemployment in the area of the property can be an indication for the economic well-being of the borrower. Note that there is no separate GDP explanatory variable, as this is not available on the county level. Additionally, an increase in unemployment rate could indicate a higher probability of a home owner suddenly defaulting (see also Smith, Sanchez, and Lawrence (1996)). The unemployment rate is quarterly, and made available by the Bureau of Labor Statistics. To construct the monthly values the quarterly values were linearly interpolated. To adjust for any look-ahead bias the series was lagged by 3 months.

Jumbo This is a dummy to indicate if the original loan value was above the conforming loan threshold of \$417,000 in effect in 2010 (Federal Housing Finance Agency (2010))[6]. Smith, Sanchez, and Lawrence (1996) find that loans above \$500,000 are significantly more likely to prepay. This is possibly due to the fact that a larger notional means that an interest rate differential will have a larger effect on the amount saved via refinancing.

Figure 3.2: Scatterplots of Credit vs. DTI

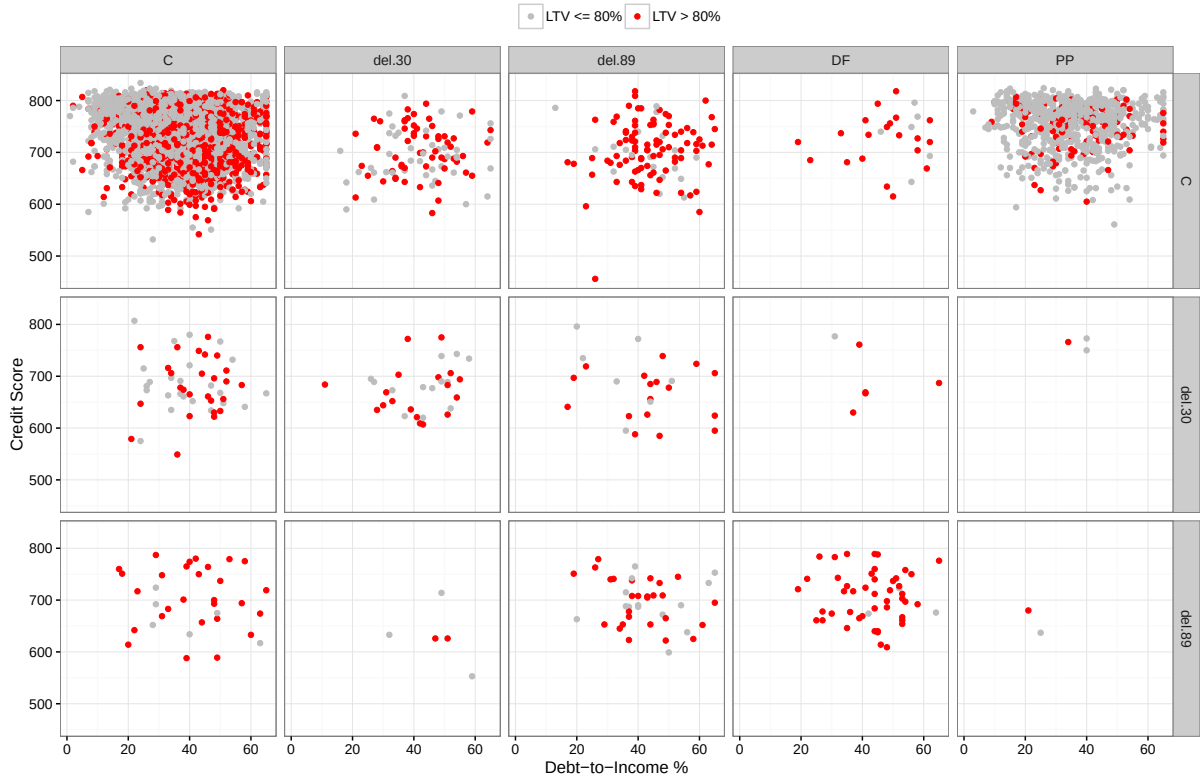


Figure shows the development of the performance of 5000 randomly sampled mortgages during 2010. The rows indicate the performance of the mortgage at the start of 2010, columns indicate the performance state at the end of 2010. For example, the 2nd row 1st column shows the credit score and debt-to-income ratios of mortgages that started the year as DELINQUENT 30-89 DAYS and ended 2010 as CURRENT.

[6]Although the conforming loan threshold can vary by county and was temporarily raised in 2008, most mortgage providers have maintained the original conforming loan definition when underwriting

4 Methodology

In section 4.1 I outline the proposed Markovian structure in mortgage performance, this will lead to the multinomial distribution of transition probabilities which I model with a multinomial logit model, discussed in section 4.2. In section 3.2 I discussed the considered explanatory variables that enter the model and their construction. The significance and implementation into the multinomial logit model are also discussed in Section 4.2.

Next I expand upon this bases by introducing segmentation in the data with the aim of achieving more homogeneous groups. I consider segmentation by geographical location, north versus south California, vintage year buckets, and a data driven k-means segmentation. Details on this can be found in Section 4.3 and 4.4. Finally with the aim of reducing the parameter uncertainty arising from the segmentation of the data, in Section 4.5 I propose a Bayesian approach to the multinomial logit model. The parameters are then estimated with the use of an adaptive Metropolis-Hastings algorithm, explained in Section 4.6.

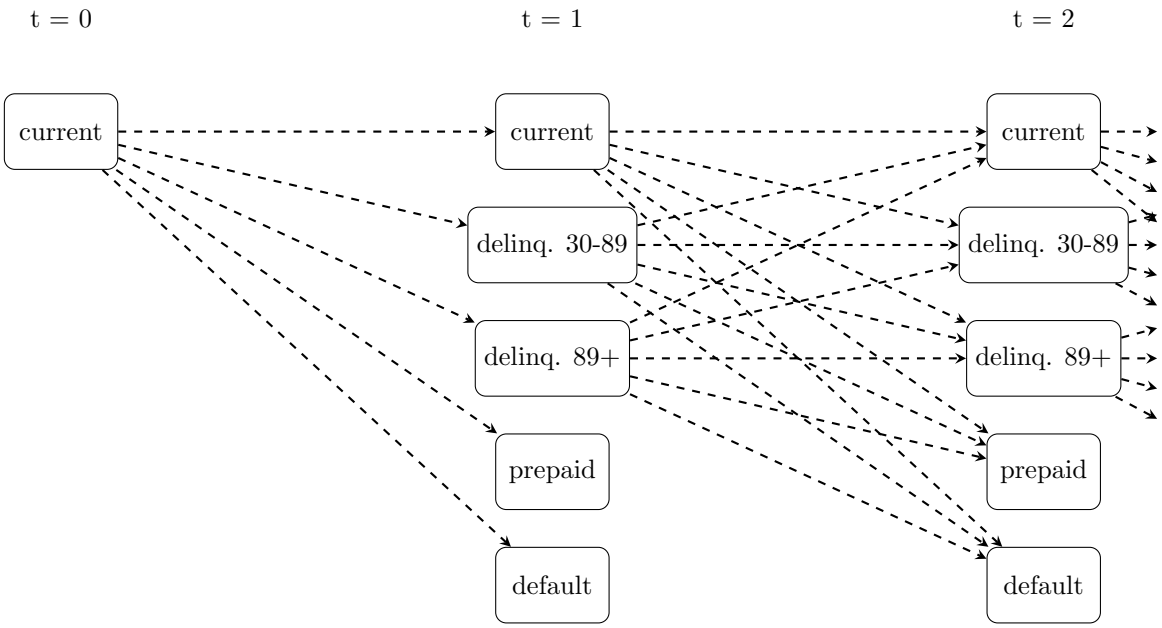
I end the methodology section with an outline of the model comparison methods in Section 5.

4.1 Markovian Structure

I follow Smith, Sanchez, and Lawrence (1996) in the general setup of modeling the performance of the mortgage loans. I start by defining a Markovian structure with non-stationary transition probabilities to model the life of a mortgage. Consider some time t and some loan m. At time t loan m can be in a select number of predefined states indicating a unique characteristic the loan can find itself in. As discussed earlier I consider 5 states: CURRENT; the loan is up to date on it's payments, DELINQUENT 30-89 DAYS; the loan is 30 to 89 days behind on payments, DELINQUENT 89+ DAYS; the loan is more than 89 days behind on its payments, DEFAULT; the loan has defaulted on its obligations, and PREPAYMENT/PAID-OFF; the loan is paid off. Mortgages that are CURRENT can also be denoted as a performing, whereas mortgages that are in DELINQUENCY 30-89 DAYS and DELINQUENCY 89+ DAYS are non-performing mortgages. Between two time periods a mortgage can transition from its state at t to a state at time $t+1$.

Modeling the evolution of a mortgage through these predefined states allows to take into account the inherent differences present in each of the states. For example, it is not unreasonable that a borrower that is behind on its monthly payments will have significantly different motivations to prepay their mortgage than a mortgage that is current with its payments. Likewise, drivers leading to a DEFAULT state from a CURRENT state will be different than those that influence the transition from a DELINQUENT 89+ DAYS state. Figure 3.2 in Section 3.2 gives an overview of these differences between performing and non-performing states. Grimshaw & Alexander (2011) decided to model 8 unique states, wherein they include more delinquency states. However, Grimshaw & Alexander (2011) model the transition probabilities themselves without any explanatory variables, concentrating on the frequency of observed transitions. As I allow my, to be estimated, transition probabilities to be driven by covariates, the inclusion of so many unique states would become impractical. As such I choose to define only the above mentioned 5 states. This allows for some differentiation in delinquency characteristics, whilst also maintaining a relative parsimonious model. In any case, in this setup the state a mortgage m is in is something that is observed. One can make use of hidden Markov models to drop this restriction, but that is beyond the scope of this research. Lastly note that the states PREPAID and DEFAULT in Figure 4.1 are absorbing nodes. Once a mortgage reaches one of these states they can never leave.

Figure 4.1: 5-state Markovian structure



A schematic overview of the development of the performance of a mortgage over time. Note that DEFAULT and PREPAID are absorbing nodes and as such do not have any out going connections.

Following the work of Sung, Soyer, & Nhan (2007) let us define $s_{m,t}$ as a random variable that takes a value in $\{1 \dots J\}$, indicating the state mortgage m is in at time t. With $j \in \{1, \dots, J\}$ being the possible states, t the time, and m the mortgage. Note that $J = 5$ in my setup and refers to the above mentioned 5 states. I maintain the general notation for convenience. I assume that the sequence $\{s_{m,0}, s_{m,1}, \dots, s_{m,T}\}$ follows a first order Markov chain such that $p(s_{m,t}|s_{m,t-1}, s_{m,t-2}, \dots) = p(s_{m,t}|s_{m,t-1})$. Less formally, the probability of being in state j at time t given that the mortgage was in state i at time $t-1$ is only dependent on state i at time $t-1$ and not its predecessors. Furthermore, let us denote $x^i_{m,j,t}$ as a binary variable indicating a transition occurring from state i at $t-1$ to state j at time t for mortgage m:

$$x^i_{m,j,t} = 1(s_{m,t} = j|s_{m,t-1} = i) \tag{4.1}$$

where $1(A)$ is the identification function. Then the vector $\boldsymbol{x}^i_{m,t} = [x^i_{m,1,t}, \dots, x^i_{m,J,t}]$ is a multinomial random variable indicating the observed transition of mortgage m at time t. The probability vector of $\boldsymbol{x}^i_{m,t}$ can be defined as $\boldsymbol{\pi}^i_{m,t} = [\pi^i_{m,1,t}, \dots, \pi^i_{m,J,t}]$ such that $\pi^i_{m,j,t} = p(s_{m,t} = j|s_{m,t-1} = i)$ and $\sum_{j=1}^{J} \pi^i_{m,j,t} = 1$. Thus defining the distribution of $\boldsymbol{x}^i_{m,t}$ as:

$$(\boldsymbol{x}^i_{m,t}|\boldsymbol{\pi}^i_{m,t}) \sim Multinomial(\boldsymbol{\pi}^i_{m,t}) \tag{4.2}$$

As can be seen each of the 5 defined states has an associated transition probability vector $\boldsymbol{\pi}^i_{m,t}$ with $i \in \{$ CURRENT, DEL. 30-89 DAYS, DEL. 89+ DAYS, DEFAULT, PREPAID$\}$. In practice the transition probabilities starting from either PREPAYMENT or DEFAULT are already known for all t and m as these are absorbing nodes. The matrix of transition probabilities from any state i to any state j for a mortgage m at time t is then:

$$\mathbf{\Pi}_{m,t} = \begin{bmatrix} \pi^1_{m,1,t} & \cdots & \pi^1_{m,J,t} \\ \vdots & \ddots & \vdots \\ \pi^J_{m,1,t} & \cdots & \pi^J_{m,J,t} \end{bmatrix} \tag{4.3}$$

I model these multinomial transition probabilities via a multinomial logit model for each of the three possible starting states (CURRENT, DELINQUENT 30-89 DAYS, DELINQUENT 89+ DAYS). This is further discussed in the next section. For ease of reading, I will drop the index denoting the month t in the remainder of the paper. As the models are estimating over 2010, and forecast 2011, the index can be dropped without ambiguity.

4.2 Mulitnomial Logit

In this section I expand on estimating the transition probabilities as noted in Equation (4.3). To incorporate the effect of explanatory variables on the transition probabilities of each Markov chain I model the transition probabilities from each state i to state j with a multinomial logistics regression. For this I follow the work of Sung, Soyer, & Nhan (2007) whom apply a time homogeneous Markov Chain to the assessment of psychiatric treatment programs.

Let us denote, as before, the transition vector $\boldsymbol{\pi}^i_m = [\pi^i_{m,1}, \ldots, \pi^i_{m,J}]$ as the transition probabilities of loan m going from state i to state j. The logit transformation of the transition probability $\pi^i_{m,j}$ can be written as:

$$\eta^i_{m,j} = \text{logit}(\pi^i_{m,j}) = \log(\frac{\pi^i_{m,j}}{\pi^i_{m,1}}) = \boldsymbol{F}_m \boldsymbol{\theta}^i_j \tag{4.4}$$

for $i, j \in \{1, \ldots, J\}$, and $\boldsymbol{F}_m$ is a $[1 \times Q]$ vector of explanatory variable values for loan m. Additionally, $\boldsymbol{\theta}^i_j$ is a $[Q \times 1]$ vector of regression parameters. The 1-st category is used as a baseline category in (4.4). The transition probability $\pi^i_{m,j}$ is then given by:

$$\pi^i_{m,j} = \frac{exp(\boldsymbol{F}_m \boldsymbol{\theta}^i_j)}{\sum_{j=1}^{J} exp(\boldsymbol{F}_m \boldsymbol{\theta}^i_j)} \tag{4.5}$$

Equation (4.4) can be written more generally as a multivariate logit transformation:

$$\boldsymbol{\eta}^i_m = \text{logit}(\boldsymbol{\pi}^i_m) = \boldsymbol{F}_m \mathbf{\Theta}^i \tag{4.6}$$

where $\boldsymbol{\eta}^i_m = [\eta^i_{m,1}, \ldots, \eta^i_{m,J}]$ a $[1 \times J]$ vector of logit transforms and $\mathbf{\Theta}^i$ a $[Q \times J]$ regression parameter matrix:

$$\mathbf{\Theta}^i = \begin{bmatrix} \theta^i_{1,1} & \cdots & \theta^i_{J,1} \\ \vdots & \ddots & \vdots \\ \theta^i_{1,Q} & \cdots & \theta^i_{J,Q} \end{bmatrix} \tag{4.7}$$

Note that the vector of regression parameters for the transition probabilities from state i to state j (denoted by $\boldsymbol{\theta}^i_j$) are represented as the j-th column of $\mathbf{\Theta}^i$. Likewise, the effect of the q-th covariate on the transitions from state i are on the q-th row of $\mathbf{\Theta}^i$. More specifically, the q-th row can be defined as $\boldsymbol{\theta}^i_q = [\theta^i_{1,q}, \ldots, \theta^i_{J,q}]$. In Section 4.5 I will assign a prior distribution to each of these rows.

I estimate the model based on mortgages that were active at the start of 2010 and observe their transitions one year later. $\boldsymbol{F}_m$ as such indicates the value of the explanatory variables at the start of 2010. $\boldsymbol{x}^i$ are then vectors of 0 and 1s indicating if the mortgage m transitioned from i to j at the end of 2010. Note that a transition to DEFAULT or PREPAID at any time during the year results in that specific mortgage also being in the respective state at the end of the year. $\boldsymbol{\Theta}^i$ can then be estimated via maximum likelihood for $i \in \{$CURRENT, DEL. 30-89 DAYS, DEL. 89+ DAYS$\}$ corresponding to each of the possible departing states. The likelihood can be derived from Equations (4.1) and (4.5):

$$p(\boldsymbol{x}^i|\boldsymbol{\Theta}^i) = \prod_{m=1}^{M}\prod_{j=1}^{J}\left(\frac{exp(\boldsymbol{F}_m\boldsymbol{\theta}_j^i)}{\sum_{j=1}^{J} exp(\boldsymbol{F}_m\boldsymbol{\theta}_j^i)}\right)^{x_{m,j}^i} \tag{4.8}$$

the log-likelihood is then:

$$\log p(\boldsymbol{x}^i|\boldsymbol{\Theta}^i) = \sum_{m=1}^{M}\sum_{j=1}^{J} x_{m,j}^i \ln\left(\frac{exp(\boldsymbol{F}_m\boldsymbol{\theta}_j^i)}{\sum_{j=1}^{J} exp(\boldsymbol{F}_m\boldsymbol{\theta}_j^i)}\right) \tag{4.9}$$

Which can be optimized to obtain $\hat{\boldsymbol{\Theta}}^i_{\text{ML}}$, the maximum likelihood estimate of the multinomial logit model. The standard errors of the coefficients can be retrieved via the inverse of the information matrix evaluated at the maximum likelihood estimate $(\mathcal{I}^{-1}(\hat{\boldsymbol{\Theta}}^i_{\text{ML}}))$.

To determine which explanatory variables are used in each of the 3 departing states multinomial logit models, I follow the approach of Smith, Sanchez, and Lawrence (1996). A general to specific approach was applied to determine which covariates are implemented in the model. A covariate remained in the model if it had a significance level of 10% or lower for any of the transition probabilities. Covariates were then removed iteratively. This process was repeated for each of the three possible starting performance states (CURRENT, DELINQUENT 30-89 DAYS, and DELINQUENT 89+ DAYS).

Table 2: Explanatory variables used

Variable	Type	Included Current	Included Del. 30-89	Included Del. 89+
Provided				
Credit Score	numerical	x	x	x
Original LTV	numerical			
Original DTI	numerical	x		
Occupancy Status	categorical	x		
Number of borrowers	categorical	x		
Channel	categorical			
First time homebuyer	categorical			
Loan purpose	categorical	x		
Constructed				
Log(loan age)	numerical	x		
Every delinquent	categorical	x		
Under water	categorical			
Time since profitable	numerical	x		
Estimated LTV	numerical	x	x	x
Interest rate differential	numerical	x	x	
County Unemployment Rate	numerical	x	x	
Jumbo	categorical	x	x	

A list of which covariates are used in each of the 3 departing state multinomial logit models.

It can also be noted that the explanatory variables are mixed between numeric and categorical. To include a categorical variable with N categories then $(N-1)$ dummy variables are created. The iterative process of determining the covariates to use was done in a frequentist manner as this process in the Bayesian framework of Section 4.5 would have taken significantly more time. This is mainly due to the sampling of the posterior to acquire the parameter estimates in the Bayesian setting.

4.3 Segmentation

As outlined in Section 4.1 I implement a Markovian structure as proposed by Smith, Sanchez, and Lawrence (1996) to model the development of mortgage performance. I expand on this model by further segmenting the entire mortgage pool into, assumed, homogeneous groups of mortgages. The choice of segmenting the entire pool of considered mortgages is motivated in part by Liang & Lin (2014) who use a two-stage approach to forecasting prepayment behavior. They make use of a random forest to define the segments and find it results in more accurate forecasts. On a more intuitive note, grouping mortgage loans by, for example, origination year (a big factor in underwriting requirements), or geographic location of the property could yield more homogeneous segments than just grouping all loans together. For example, borrowers of mortgages that originate in the south of California near San Diego and the Mexican border are culturally different from those farther north near San Francisco. It can be that between these geographical locations the drivers behind the choices of a borrower can be different. Likewise, loans that originated prior to the financial crisis can have been subjected to less stringent underwriting requirements than those post crisis. These differences create heterogeneity in the mortgage pool, and cause mortgages in different segments to react differently to the covariates. The idea is to try to split the larger pool of mortgages in the segments that are more homogeneous.

Let us take the entire pool of considered mortgages and divide these into L segments (groups) each consisting of assumed homogeneous mortgage loans. These segments will be denoted by $\mathcal{L}$ with $\mathcal{L} \in \{1 \dots L\}$. The main idea is that mortgages in the same segment $\mathcal{L}$ are assumed to have the same relation to the underlying explanatory variables. The Markovian structure, as discussed in Section 4.1 is then applied to each segment separately. Such that for each starting performance state CURRENT, DELINQUENT 30-89 DAYS, and DELINQUENT 89+ DAYS there will be L multinomial logit models to estimate. I consider 3 segmentation schemes. 2 intuitive, by geographical location and by vintage year, and 1 data driven, using a k-means algorithm. Segmentation by geographical location of the underlying property will be mainly discussed and investigated during the main body of this paper (e.g. when discussing parameter estimation and prior sensitivities). Geographical and vintage segmentation are further discussed and motivated in this section. In Section 4.4 I outline the use of a k-means algorithm on the estimation sample to segment the mortgages, and further discuss the characteristics of the resulted groups.

Table 3: Geographical segmentation statistics

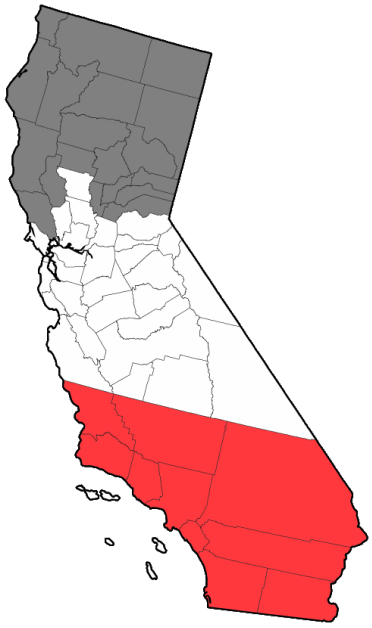
	Geographic			**Vintage**				**K-means**			
	North	Central	South	1999-2002	2003-2006	2007	2009-2011	Normal	Low risk	Mature	High risk
Mortgages	1020	6069	9339	715	6639	1784	7290	5727	4761	4256	1684

Table shows the number of mortgages in each segment for each segmentation scheme.

For geographical segmentation I have chosen to create 3 segments; *north*, *central*, and *south* California. An overview of this segmentation can be seen in Figure 4.2. I use the 3 digit postal code available in the

data-set to map each mortgage to one of the three segments. The precise distribution of the postal codes over the 3 segments can be found in Appendix A.2. In short, the segmentation follows the known cultural division between North California and South California. The *south* segment includes Los Angeles and San Diego and the *central* segment includes the San Francisco and Sacramento areas. Lastly the *north* segment are the rural areas north of San Fransisco, and by far the smallest of the 3 segments in terms of population.

Figure 4.2: Geographical segmentation of California



The three colors indicate the *North*, *Central*, and *South* segments of California.

In Table 4 we can see the differences between these three regions. First of all the *south* region tends to have higher delinquency and default rates. Where as the rural *north* region has higher unemployment, but also cheaper homes. Of note is that in the *north* the unemployment rates are higher en the prepayment rates are lower, but this is not reflected in higher default rates. This could indicate that underwriting practices are more stringent in this area than compared to the *central* or *south*.

Table 4: Geographical segmentation statistics

									Start 2010 as		*In 2010*	
	Avg. unemp	DTI	Est. LTV	Credit score	Jumbo	Int. diff	Loan age	Ever del.	Del. 30-89	Del. 89+	df	pp
All	9,2%	36,9%	72,1%	739	48,2%	0,21	36	14,5%	3,1%	4,3%	1,8%	18,7%
Geographic												
Central	9,1%	36,3%	74,0%	744	50,1%	0,18	36	12,5%	2,4%	3,5%	1,7%	20,5%
North	10,2%	36,3%	70,9%	738	29,5%	0,28	41	12,4%	3,1%	2,8%	1,6%	15,2%
South	9,1%	37,3%	71,1%	736	49,1%	0,23	36	16,0%	3,5%	4,9%	1,9%	17,9%

Table shows the average values for the mortgages in each geographical segment. Values are dated at the start of 2010, with the exception of the default and prepayment rates, which are rates for the entire year. statistics for the different k-means segments.

The segmentation by vintage is done in four groups of origination years. 1999-2002, 2003-2006, 2007 and 2008-2011 are the considered groups. As such it will consist of $L = 4$ segments. Segmentation by origination year is motivated by the idea that laws pertaining to regulations and requirements for taking

out mortgages can change over time. A mortgage borrower who acquired a loan during a time of lenient lending practices will find it difficult to refinance his or her mortgage in a time when these laws have become stricter. This would cause the mortgage borrower to perceive a higher cost to refinance than loans that originated in other periods. Grimshaw & Alexander (2011) choose to segment by origination year (actually even specific quarter of each year) and find that this improves upon forecasting accuracy. Note that this segmentation also captures, in addition to the possible differences in underwriting practices, the effect of the mortgage age. Which has also been shown to be a significant factor in mortgage borrower behaviour (see for example Schwartz & Torous (1989) or Kang & Zenios (1992)). The groups are however, in general, spanning a long enough period to motivate the use of the covariate *log(loan age + 1*. The choice of the above defined vintage segments was made in an effort to capture loans that originated in pre-crisis and those that originated post-crisis. *2007* was also denoted as a separate segment as this was the period wherein the crisis took hold. Additionally the segment 1999-2002 is added as this consists of mortgages that were acquired early on and are now very matured in comparison to the other vintage segments.

Table 5 below shows some statistics for these segments. It can be seen that mortgages that originated in the *crisis* and *pre-crisis* period have higher Debt-to-Income ratios as well as higher Loan-to-Value ratios. Credit ratings also tend to be a bit higher post-crisis versus pre-crisis. Additionally, *crisis* and *pre-crisis* mortgages show significantly higher delinquency and default rates. Also note that the estimated Loan-to-Value ratio for *early* mortgages is low due to the long period wherein these loans have been repaying on notional.

Table 5: Vintage segmentation statistics

Vintage	Avg. unemp	DTI	Est. LTV	Credit score	Jumbo	Int. diff	Loan age	Ever del.	*Start 2010 as* del. 30-89	del. 89+	*In 2010* df	pp
crisis	9,7%	39,5%	94,2%	723	51%	0,69	30	25,5%	6,2%	11,1%	4,4%	13,4%
early	9,2%	33,8%	41,9%	728	12%	1,16	99	21,4%	1,1%	1,6%	0,3%	18,6%
postcrisis	8,8%	36,7%	68,8%	756	56%	-0,07	11	6,6%	1,9%	2,1%	1,1%	21,3%
precrisis	9,4%	36,8%	73,3%	726	43%	0,30	59	19,5%	3,8%	5,1%	2,1%	17,2%

Table shows the average values for the mortgages in each vintage segment. The comments of Table 4 also apply.

The segmentation via the k-mean algorithm and the resulting segments (and their characteristics) are further discussed in Section 4.4. On a general note, the parameter matrix that is to be estimated for each departing state i, $\boldsymbol{\Theta}^i$, is to be separately estimated for each $\mathcal{L} \in \{1, \ldots, L\}$. Such that the notation in Equation (4.7) becomes:

$$\boldsymbol{\Theta}^{i,\mathcal{L}} = \begin{bmatrix} \theta_{1,1}^{i,\mathcal{L}} & \cdots & \theta_{J,1}^{i,\mathcal{L}} \\ \vdots & \ddots & \vdots \\ \theta_{1,Q}^{i,\mathcal{L}} & \cdots & \theta_{J,Q}^{i,\mathcal{L}} \end{bmatrix} \tag{4.10}$$

Where $\boldsymbol{\theta_q^{i,\mathcal{L}}} = [\theta_{1,q}^{i,\mathcal{L}}, \ldots, \theta_{J,q}^{i,\mathcal{L}}]$ is the vector of parameters representing the transition probability sensitivities to covariate q, for departing state i in segment $\mathcal{L}$. These parameter vectors will have a prior distribution applied to them, which is further discussed in Section 4.5.

It should be noted that we saw in Tables 4 and 5 that the different segments contain mortgages with clearly different characteristics. One way to incorporate this observation would be to include a dummy variable to account for the segment. For example, a dummy variable to indicate if a mortgage is in *south* California or not. The interpretation of this dummy variable in the multinomial logit framework is that the

geographical location will be modeled as a risk factor, where we have already seen that the geographical location of a mortgage determines its broader characteristics. In other words, just adding a dummy for the various segments will not allow for mortgages in different segments to react differently to the covariates. Of course a separate multinomial logit model for each segment brings with it different obstacles which will be tackled in Section 4.5.

4.4 K-means optimal segmentation

In contrast to segmenting by geographical location or vintage year of the mortgage, I propose the use of a k-means algorithm to divide the mortgages into more homogeneous groups. Instead of choosing the segments buy assumed differences I let the data define optimal sub groups of the mortgages. K-means clustering was first called such by MacQueen (1967), although the technique goes further back. For a detailed outline of the clustering technique see Hartigan (1975), here I will shortly discuss the basics.

Let's start with a set of multivariate elements x_i, with $i \in \{1, \dots, N\}$ and each observation x_i is of dimension $[1 \times q]$. Later we will see that I choose $q = 4$, and as such have 4 variables that I use to segment the data. The idea is to partition this set of N multivariate elements into K clusters, where each cluster has a center Q_k such that some cost measure c is minimized:

$$c = \sum_{k=1}^{K} \sum_{i \in N_k} d(x_i, Q_k) \tag{4.11}$$

Where $d(\dots)$ is some distance measure, for example Euclidean ($\sqrt{\sum_q (x_{i,q} - Q_{k,q})^2}$). N_k are the set of elements that are in cluster k, and Q_k is the center of cluster k. The center per cluster, Q_k is often taken to be the mean value per dimension. In general the k-means clustering algorithms follows the following steps:

1. Assign each element a random cluster
2. Calculate the cluster center of each cluster as the mean value for each dimension
3. For each element, find the closest cluster center and assign the element to this cluster if it is not already a member
4. Repeat (2) and (3) until no more elements are switching clusters and the algorithm has converged

As the initial clustering is assigned at random, the algorithm can be sensitive to starting conditions. As such it is not uncommon to run the algorithm multiple times and take the clustering at convergence of the run where c is lowest. I implement the approach of Hartigan & Wong (1979) which is known for its relative fast convergence. Their algorithm minimizes in the in cluster sum of squared Euclidean distance, so that $d(\dots)$ in Equation 4.11 then becomes: $d_i = \sum_q (x_{i,q} - Q_{k,q})^2$, with q being the dimensions of the elements x_i and Q_k the center element of cluster k. To account for converging to a clustering far off from the global minimizing set, due to initial conditions, I run the algorithm 100 times to ensure a robust solution.

I make use of the variables *Estimated loan-to-value*, *Loan age*, *County unemployment rate*, and *Interest rate differential* as the dimensions over which the clustering takes place. These 4 variables can be interpreted to proxy the following 4 characteristics:

1. Willingness to pay (*Estimated loan-to-value*). The relative value of the underlying property to the outstanding notional, or lack thereof, can (de)motivate a borrower to repay their loan.
2. Willingness to pre-pay (*Interest rate differential*). If a borrower can refinance the mortgage with a lower rate they have a financial incentive to do so.

3. Ability to pay (*County unemployment rate*). As a proxy for economic well-being of the area, a higher unemployment rate can indicate the inability to repay.
4. Ability to pre-pay (*Loan age*). Consider a newly originated mortgage. It is unlikely that a mortgage will pre-pay in the first few months because if the borrower had this ability they would most likely have taken a larger mortgage.

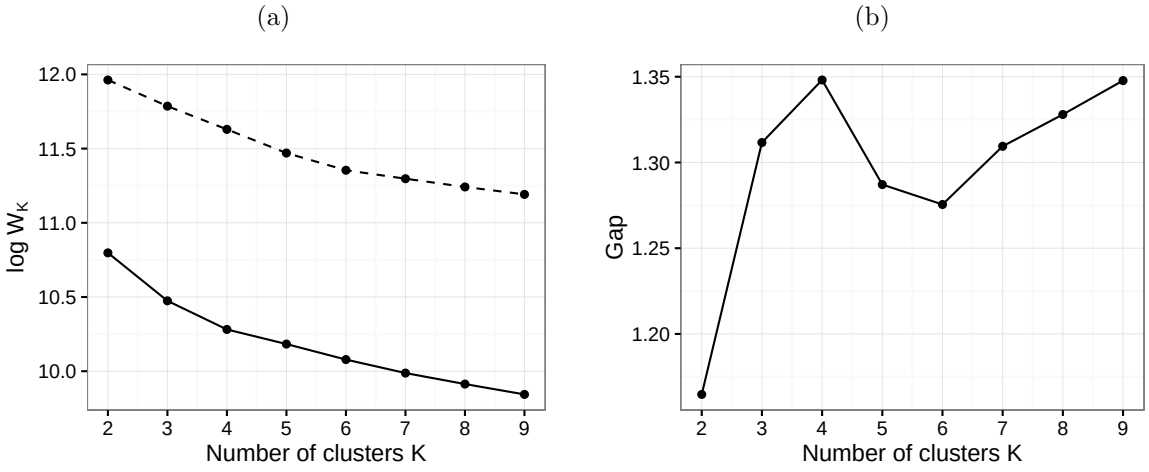
Additionally, these 4 variables are all numeric, allowing for easy valuation of the distance function $d(\dots)$. Finally, these 4 variables are updated monthly and reflect real time characteristics of the mortgage. This in contrast to, for example, *Credit score*, which is only known at origination of the mortgage. The goal is to achieve a set of cluster of mortgages, such that mortgages in the same clusters behave similarly. Each of the 4 above discussed variables is demeaned and scaled by their respective standard deviations prior to applying the K-means algorithm to determine the clusters.

When applying a K-means algorithm to cluster data a choice must be made for the number of clusters K to use. In fact a major challenge in cluster analysis is determining the optimal number of clusters in a data-set. Various methods have been proposed to determine the optimal number of clusters, see Milligan & Cooper (1985) for a comprehensive discussion. An intuitive first approach is to look at the within cluster sum of squared distances to the cluster centers:

$$D_k = \sum_{x_i \in N_k} \sum_{q} (x_{i,q} - Q_{k,q})^2 \tag{4.12}$$

With N_k being the set of elements belonging in cluster k. Then $W_K = \sum_{k=1}^{K} D_k$ is the pooled sum of squared distances to the cluster centers. Plotting W_K for different values of K can then give some insight on the optimal number of clusters.

Figure 4.3: W_K and Gap as a function of number of clusters K



(**a**) The log pooled sum of squared distances to the cluster centers versus the choice of K clusters. Solid line is $\log W_K$ of the mortgage data. Dashed line is the $\log W_K$ of the baseline simulated data that has no inherent clustering. (**b**) The Gap curve as a measure of distance between the $\log W_K$ of the mortgage data versus the $\log W_K$ of the baseline simulated data.

The solid line in Figure 4.3(a) shows the development of the sum of squared distances as the number of clusters goes from 1 to 9. As the number of clusters increases the sum of squared distances decreases, as is to be expected. One measure to determine when an optimal number of clusters is reached is the "Elbow"

method. This is where one takes the break point, or "Elbow", in the plot of W_K versus K to be the optimal number of clusters. Sadly, often it is the case that the "Elbow" point cannot clearly be identified, as is the case in Figure 4.3(a).

I therefore follow the work of Tibshirani *et al.* (2001) who propose the use of a Gap statistic. The idea of Tibshirani *et al.* (2001) is to standardize $logW_K$ in Figure 4.3(a) with some base line reference. As reference they propose the expected value $E_{n0}[\log(W_K^*)]$ under a baseline reference distribution of the data. In other words, if the data under consideration exhibits clustering, than it would be expected that the log sum of squared distances will decrease stronger with K when compared to simulated data with the same number of observations without this underlying cluster structure. Defining:

$$\text{Gap}(K) = E_{n0}[\log(W_K^*)] - \log(W_K) \tag{4.13}$$

Then the optimal number of clusters is the first value of K where $Gap(K) > Gap(K+1)$. Tibshirani *et al.* (2001) consider two approaches to constructing the reference distribution of the data. If n is the number of observations in the considered data set:

1. Generate n observations uniformly from the box spanned by the range of the dimensions of the data
2. Generate n observations uniformly from the box spanned by the principle components of the data.

I make use of the former to generate the reference values of $\log(W_K)$. To this end I follow Tibshirani *et al.* (2001) and implement a Monte Carlo simulation and simulate $B = 100$ reference data-sets. The reference data-sets are created by sampling for each dimension (i.e. 4) uniformly between the minimum and maximum value observed for that dimension. As such, a sampled observation is a vector of $[1 \times 4]$, where each column is a sampled value from *Estimated loan-to-value*, *Interest rate differential*, *County unemployment rate* and *Loan age* respectively. The same number of observations as in the original data-set under consideration are sampled. For each $k \in \{1, \ldots, K\}$ under consideration do:

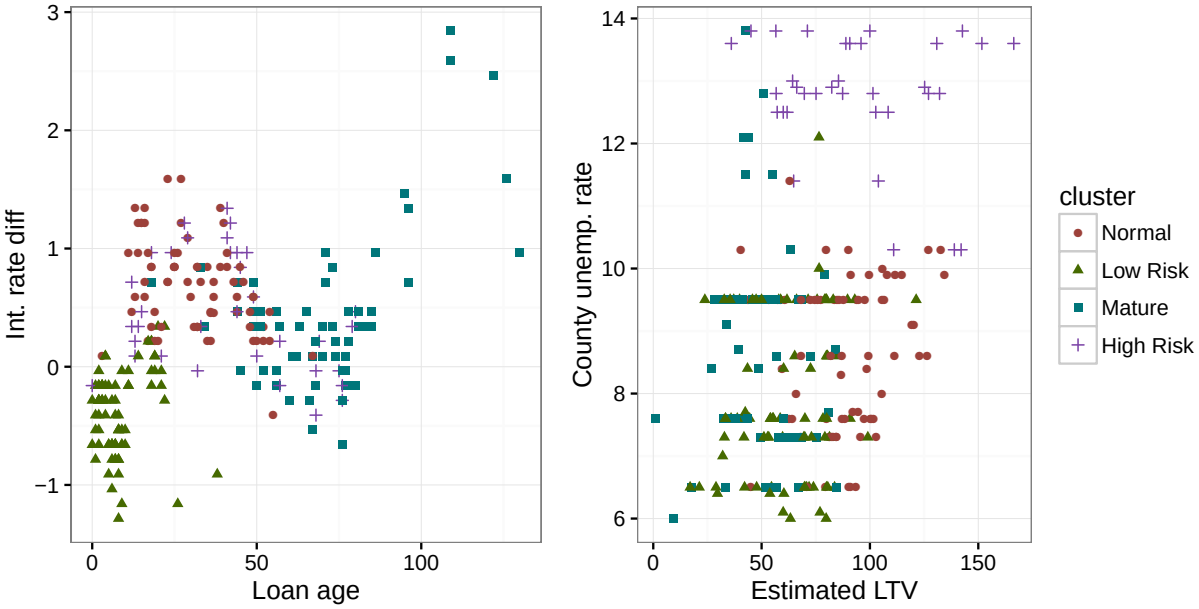
1. Apply the K-means algorithm of Hartigan & Wong (1979) to the mortgage data to acquire the allocation of elements to clusters
2. Calculate $\log(W_K)$
3. For each $b \in 1 : B$:
   (a) Simulate a data-set distributed uniformly over the observed minimum and maximum values of the 4 dimensions.
   (b) Apply the K-means algorithm to the simulated data and acquire the allocation of elements to clusters
   (c) Calculate $\log(W_{K,b}^*)$
4. Calculate $E_{n0}[\log(W_K^*)] = \frac{1}{B}\sum_{b=1}^{B}\log(W_{K,b}^*)$
5. Calculate $\text{Gap}(K) = E_{n0}[\log(W_K^*)] - \log(W_K)$

Choose the number of clusters as the first value of K where $Gap(K) > Gap(K+1)$ [7]. Figure 4.3(b) shows the result, and it can be seen that the optimal choice falls on $K = 4$ clusters.

[7] For simplicity I left out a term correcting for Monte Carlo error. For completeness, we would want to choose the first K where $Gap(K) > Gap(K+1) - s_{K+1}$ with $s_K = sd_K\sqrt{1+\frac{1}{B}}$ and $sd_K = \sqrt{\frac{1}{B}\sum_b(\log(W_K^*) - \bar{l})^2}$, lastly $\bar{l} = \frac{1}{B}\sum_b \log(W_K^*)$. Due to the number of observations, and my choice for $B = 100$ the Monte Carlo error is so small that this does not change the choice for optimal K.

The mortgage data is segmented into four different clusters as allocated by the K-means algorithm. To get a feel for how this partitions the mortgages Figure 4.4 plots the four variables used for clustering versus each other. Additionally, Table 6 shows the statistics for the K-means segmentation as was also provided for the geographic and vintage segmentation schemes earlier.

Figure 4.4: Clusters



250 randomly sampled mortgages. Different shapes and colours indicate different clusters.

Table 6: K-means segmentation statistics

K-means	Avg. unemp	DTI	Est. LTV	Credit score	Jumbo	Int. diff	Loan age	Ever del.	*Start 2010 as* Del. 30-89	*Start 2010 as* Del. 89+	*In 2010* df	*In 2010* pp
Normal	8,8%	40,3%	88,9%	727	47,9%	0,60	32	20,7%	5,3%	8,0%	3,3%	17,3%
Low risk	8,5%	34,5%	60,4%	765	66,1%	-0,49	8	3,1%	0,5%	0,7%	0,3%	21,0%
Mature	8,8%	34,3%	48,2%	734	41,5%	0,35	72	14,8%	2,1%	1,1%	0,2%	22,5%
High risk	13,0%	38,8%	109,3%	718	16,0%	0,52	41	24,5%	5,6%	9,8%	4,8%	7,0%

Table shows the average values for the mortgages in each k-means segment. The comments of Table 4 also apply.

From Figure 4.4 and Table 6 some characteristics can be deduced and as such interpret the clusters as follows:

1. **Normal**: Mortgages that are young to average of *loan age*, positive interest rate diff., low to average *county unemployment*, and average to high *estimated LTV*. This implies young to medium term normal mortgages, that have not had as long as their more mature counterparts to repay part of the notional.
2. **Low risk**: Young mortgages with below average *estimated LTV* and low *county unemployment*. These are relatively low risk as a significant amount was paid down on the mortgage.
3. **Mature**: Old mortgages with low *estimated LTV* and low *county unemployment rate*. These are low risk mature mortgages, that have had ample time to repay on the notional.

4. **High risk**: Average *loan age* mortgages with high *county unemployment rates*, high *estimated LTV* and average to high *interest rate differential*. This indicates risky mortgages.

As can be seen the K-means clustering resulted in four segments, where each segment can be uniquely characterized to imply a certain type of mortgage.

4.5 Bayesian Multinomial Logit

The proposed segmentation into more homogeneous groups as discussed in the previous sections will lead to groups with fewer observations. As can be seen in Table 3 some segments have significantly less observations than other segments. If we also consider the number of covariates that enter the model, see Table 2, it can quickly become problematic to estimate a multinomial logit model for each of the segments (and for each of the departing performance states). To tackle this deficiency I propose the use of a Bayesian approach to the multinomial logit model.

I am interested in estimating $\boldsymbol{\theta}_q^{i,\mathcal{L}}$ for each departing state i and each segment $\mathcal{L}$. More precisely, I am interested in $\boldsymbol{\theta}_q^{i,\mathcal{L}}|\boldsymbol{F}^{i,\mathcal{L}}$. Using Bayes Theorem (Bayes (1763)) it can be shown:

$$p(\boldsymbol{\theta}_q^{i,\mathcal{L}}|\boldsymbol{S}^{i,\mathcal{L}}) \propto p(\boldsymbol{S}^{i,\mathcal{L}}|\boldsymbol{\theta}_q^{i,\mathcal{L}})p(\boldsymbol{\theta}_q^{i,\mathcal{L}}) \tag{4.14}$$

Where $\boldsymbol{S}^{i,\mathcal{L}} = \{s_{1,1}, \ldots, s_{m,t}, \ldots, s_{M,t}\}$ the chains of states for all m in segment $\mathcal{L}$ and departing state i (i.e. the data).

Sung, Soyer, & Nhan (2007) also implement a Bayesian framework for the multinomial logit model. Although in their case it is applied to clinical studies. I broadly follow their approach in assigning a multivariate normal prior to each row of Eq. (4.10):

$$\boldsymbol{\theta}_q^{i,\mathcal{L}} \sim MVN(\boldsymbol{\mu}_q^{i,\mathcal{L}}, \boldsymbol{W}_q^{i,\mathcal{L}}) \tag{4.15}$$

with $\boldsymbol{\mu}_q^{i,\mathcal{L}}$ a $[J \times 1]$ mean vector and $\boldsymbol{W}_q^{i,\mathcal{L}}$ a $[J \times J]$ covariance matrix. $\boldsymbol{\mu}$ and $\boldsymbol{W}$ are determined by the multinomial logit model on the aggregated segments of each departing state i. Such that:

$$\boldsymbol{\theta}_q^{i,\mathcal{L}} \sim MVN(\hat{\boldsymbol{\theta}}_q^{i}, g\hat{\boldsymbol{\Sigma}}_q^{i}) \tag{4.16}$$

Where $\hat{\boldsymbol{\theta}}_q^{i}$ is the multinomial logit model estimate of parameter vector q for departing state i as defined in Section 4.2. Likewise, $\hat{\boldsymbol{\Sigma}}_q^{i}$ is a diagonal matrix with the squared standard errors, of the respective multinomial logit model of departing state i, on its diagonal. Lastly g is a prior hyper parameter, that is determined by in-sample forecast accuracy. A random selection of $\frac{1}{3}$ of the mortgages in the estimation sample is reserved for calibrating g. The model is estimated over the remaining $\frac{2}{3}$ of the sample for $g \in \{0.01, 0.1, 1, 2, 5, 10, 100\}$ and used to forecast the withheld $\frac{1}{3}$ sample. The value of g that yields the best in-sample results is then used for the out-of-sample forecast.

The hierarchical setup for the Bayesian multinomial logit regression for the transition probabilities can thus be summarized as:

$$\begin{aligned} \boldsymbol{x}_m^{i,\mathcal{L}}|\boldsymbol{\pi}_m^{i,\mathcal{L}} &\sim \text{Multinomial}(\boldsymbol{\pi}_m^{i,\mathcal{L}}, 1), \\ \eta_{m,j}^{i,\mathcal{L}} &= \text{logit}(\pi_{m,j}^{i,\mathcal{L}}) = \boldsymbol{F}_m\boldsymbol{\theta}_j^{i,\mathcal{L}}, \\ \boldsymbol{\theta}_q^{i,\mathcal{L}} &\sim MVN(\hat{\boldsymbol{\theta}}_q^{i}, g\hat{\boldsymbol{\Sigma}}_q^{i}) \end{aligned} \tag{4.17}$$

4.5.1 Posterior Analysis

To analyze the hierarchical setup of Equation (4.17), and more importantly, the posterior distribution of the parameters $\boldsymbol{\Theta}^{i,\mathcal{L}}$ I continue to follow Sung, Soyer, & Nhan (2007). The joint posterior distribution of the setup in Equation (4.17) is proportional to:

$$p(\boldsymbol{\Theta}^{i,\mathcal{L}}|\boldsymbol{S}^{i,\mathcal{L}}) \propto \prod_{m\in M^{\mathcal{L}}}\prod_{i=1}^{J}\Big[p(\boldsymbol{x}^{i}_{m,j}|\boldsymbol{\Theta}^{i,\mathcal{L}})\Big]\prod_{q=1}^{Q}p(\boldsymbol{\theta}^{i,\mathcal{L}}_{q}|\boldsymbol{\mu}^{i,\mathcal{L}}_{q},\boldsymbol{W}^{i,\mathcal{L}}_{q}) \tag{4.18}$$

where $M^{\mathcal{L}}$ indicates all mortgages m in segment $\mathcal{L}$ and $S^{i,\mathcal{L}}$ the observed transition chains. The joint posterior distribution of (4.18) has no clear analytical solution so the posterior will need to be simulated. Rewriting Equation (4.18) and substituting in the multivariate normal probability density function in:

$$\prod_{m\in M^{\mathcal{L}}}\prod_{j=1}^{J}\left(\frac{exp(\boldsymbol{F}_{m}\boldsymbol{\theta}^{i}_{j})}{\sum_{j=1}^{J}exp(\boldsymbol{F}_{m}\boldsymbol{\theta}^{i}_{j})}\right)^{x^{i}_{m,j}} exp\{-\frac{1}{2}\sum_{q=1}^{Q}(\boldsymbol{\theta}^{i,\mathcal{L}}_{q}-\boldsymbol{\mu}^{i}_{q})'\boldsymbol{W}^{i\,-1}_{q}(\boldsymbol{\theta}^{i,\mathcal{L}}_{q}-\boldsymbol{\mu}^{i}_{q})\} \tag{4.19}$$

which Sung, Soyer, & Nhan (2007) rightfully denote as not being from a known density. Sampling from the posterior distribution of Equation (4.19) is the main hurdle in analyzing Bayesian multinomial logit models. This is due to the analytically inconvenient form of the models' likelihood function, as opposed to probit models where the latent-variable approach of Albert & Chib (1993) allows for easier sampling. Several approaches have been proposed in the literature. Polson *et al.* (2013), for example, propose a data-augmentation strategy based on Pólya-Gamma distributions. I however make use of a Metropolis-Hastings sampling algorithm outlined by Hastings (1970) with a random-walk kernel for its ease of implementation and general familiarity. I apply some adaptive measures in the Metropolis-Hastings technique to assure efficient sampling.

4.6 Adaptive Metropolis-Hastings algorithm

In outlining the adaptive Metropolis-Hastings algorithm I will use a general notation not directly related to notation in other sections for ease of discussion. The general idea behind this algorithm is given a sampled value $\boldsymbol{X}^{\{n\}}$ from the target distribution, generate a candidate sample, $\boldsymbol{Y}^{\{n+1\}}$, from a proposal density $q(\boldsymbol{Y}^{\{n+1\}}|\boldsymbol{X}^{\{n\}},\dots)$, then accept this proposal with a probability α defined as[8]:

$$\alpha(\boldsymbol{X}^{\{n\}},\boldsymbol{Y}^{\{n+1\}}) = \min\left\{\frac{f(\boldsymbol{Y}^{\{n+1\}}|\dots)}{f(\boldsymbol{X}^{\{n\}}|\dots)},1\right\} \tag{4.20}$$

Where $\boldsymbol{X}^{\{.\}}$ and $\boldsymbol{Y}^{\{.\}}$ are both d dimensional vectors containing the parameters to be sampled. $f(\dots)$ is the target distribution. Given the shape of the prior and the ease of simulating, I make use of a random walk multivariate normal distribution as proposal density:

$$q(\boldsymbol{Y}^{\{n+1\}}|\boldsymbol{X}^{\{n\}},\dots) = \begin{cases} \text{MVN}(\boldsymbol{X}^{\{n\}},\sigma^{\{n+1\}}\Sigma^{\{n+1\}}) & \text{if } n+1 \leq n^{*} \\ \\ \text{MVN}(\boldsymbol{X}^{\{n\}},\sigma^{\{n^{*}\}}\Sigma^{\{n^{*}\}}) & \text{else} \end{cases} \tag{4.21}$$

Where the density is centered at the accepted value $\boldsymbol{X}^{\{n\}}$. $\sigma^{\{.\}}$ and $\Sigma^{\{.\}}$ are iteratively updating parameters

[8] α defined here is the simplification in case of random walk sampling.

governing the covariance matrix of the proposal density. The aim of adapting the covariance matrix of the multivariate normal proposal density is to achieve a certain level of desired acceptance rate. This updating is only done in the burn-in phase of the sampling, where n^* indicates the number of burn-in samples.

The use of an adaptive Metropolis-Hastings algorithm is motivated by the fact that $i \times \mathcal{L}$ separate models will have to be estimated. To circumvent the process of having to manually determine the optimal parameters governing the proposal density the choice for an automatic adapting variant was made. The updating procedure follows that of Atchadé & Rosenthal (2005) and Haario *et al.* (2001). The covariance matrix $\Sigma^{\{.\}}$ of the proposal density is determined by the history of the sampled values:

$$\Sigma^{\{n+1\}} = \begin{cases} \Sigma^{\{0\}} & \text{if } n+1 \leq n_0 \\ \text{cov}(\boldsymbol{X}^{\{0\}}, \ldots, \boldsymbol{X}^{\{n\}}) & \text{else} \end{cases} \tag{4.22}$$

Where $\Sigma^{\{0\}}$ is some initial covariance matrix and n_0 is the minimum number of samples before starting to adapt. Haario *et al.* (2001) proof the ergodicity of this adaptive procedure. Atchadé & Rosenthal (2005) expand on this adaptive procedure by allowing scaling of the proposal density via $\sigma^{\{.\}}$. I implement a slightly modified version of the adaptive algorithm of Atchadé & Rosenthal (2005) by scaling the proposal density multiplicatively, instead of additive.

$$\sigma^{\{n+1\}} = \begin{cases} p(\sigma^{\{n\}})(1-\gamma) & \text{if } \bar{\alpha}(\ldots) \leq \bar{\tau} \\ p(\sigma^{\{n\}})(1+\gamma) & \text{else} \end{cases} \tag{4.23}$$

Where $\bar{\tau}$ is the target acceptance rate, $\bar{\alpha}(\ldots)$ is the average acceptance probability for the last 100 proposals, γ is a tuning parameter governing how quickly the adapting takes place, and the function $p(.)$ is a projection function limiting the value of $\sigma^{\{n+1\}}$ to valid values (i.e. positive non-zero). To warrant that the steady state distribution of the Markov chain coincide with the targeted posterior distribution, I only apply the adaptive nature of the algorithm in the burn-in phase. The sampling algorithm then becomes:

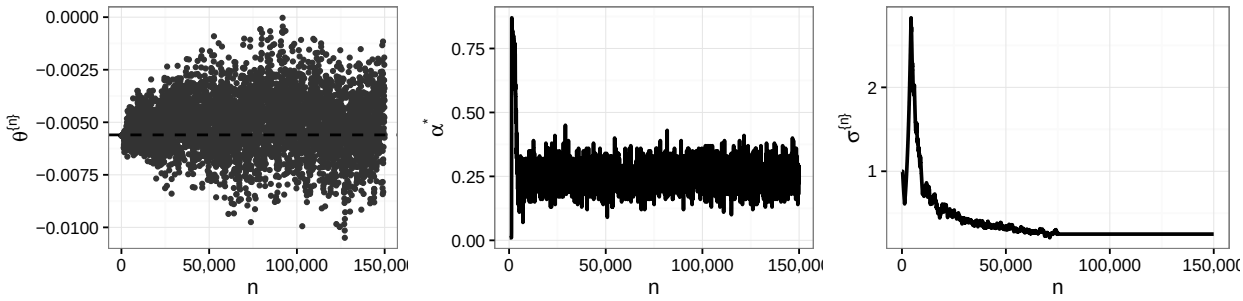
Given some sampled value $\boldsymbol{X}^{\{n\}}$ from $f(\boldsymbol{X}^{\{n\}}|\ldots)$:

1. Sample a candidate $\boldsymbol{Y}^{\{n+1\}} \sim \text{MVN}(\boldsymbol{X}^{\{n\}}, \sigma^{\{n\}}\Sigma^{\{n\}})$ and sample $U \sim \text{U}(0,1)$
2. Calculate $\alpha(\boldsymbol{X}^{\{n\}}, \boldsymbol{Y}^{\{n+1\}}) = \min\left\{\frac{f(\boldsymbol{Y}^{\{n+1\}}|\ldots)}{f(\boldsymbol{X}^{\{n\}}|\ldots)}, 1\right\}$
3. If $U \leq \alpha(\boldsymbol{X}^{\{n\}}, \boldsymbol{Y}^{\{n+1\}})$ then set $\boldsymbol{X}^{\{n+1\}} = \boldsymbol{Y}^{\{n+1\}}$. Otherwise set $\boldsymbol{X}^{\{n+1\}} = \boldsymbol{X}^{\{n\}}$
4. If $n \leq n^*$ then:
   (a) Set $\Sigma^{\{n+1\}} = \text{cov}(\boldsymbol{X}^{\{0\}}, \ldots, \boldsymbol{X}^{\{n\}})$
   (b) Set $\sigma^{\{n+1\}} = p(\sigma^{\{n\}})(1-\gamma)$ if $\bar{\alpha}(\ldots) \leq \bar{\tau}$
   Else $\sigma^{\{n+1\}} = p(\sigma^{\{n\}})(1+\gamma)$
5. Repeat from (1)

I implement the tuning parameters $\bar{\tau} = 0.25$, $\gamma = 0.01$, and $p(\sigma^{\{n+1\}}) = \sigma^{\{n+1\}}$ if $\sigma^{\{n+1\}} \geq 0.01$ and $p(\sigma^{\{n+1\}}) = \sigma^{\{n\}}$ otherwise. To summarize the procedure, I target an acceptance rate of 25% in the samples from the proposal density. To do this I constantly update the covariance matrix of the multivariate normal proposal density. This updating is done by comparing the acceptance rate for the previous 100 samples, $\bar{\alpha}$,

with the desired acceptance rate of $\tau = 25\%$. Adjustments to the covariance matrix are then based on the covariance of the sampled parameters and a scaling factor of (1 ± 0.01).

Figure 4.5: Adaptive MCMC output



$\theta^{\{n\}}$ are the generated values of $\theta^{i,\mathcal{L}}_{j,q}$ where $i =$ DELINQUENCY 30-89 DAYS, $\mathcal{L}$ is the CENTRAL geographical segmentation, j is a transition to the CURRENT performance state, and q is the explanatory variable *credit score*. The dashed line is the respective multinomial logit estimate for the corresponding parameter estimated over all segments, not just CENTRAL segment. α^* is the moving average acceptance rate of the algorithm with a window of 100 samples. $\sigma^{\{n\}}$ is value of the adapting scaling parameter. n^* (burn-in samples) was set to $75,000$ and the prior scaling (see Equation (4.16)) set to $g = 1$. Shown observations have been thinned by a factor of 40 for ease of plotting.

Figure 4.5 shows the simulation path of one parameter generated via the adaptive Metropolis-Hastings algorithm. As can be see the updating scaling parameter $\sigma^{\{n\}}$ first increases, allowing the algorithm to take larger steps when in a relatively flat area of the posterior distribution. As the algorithm converges, it also becomes steeper, requiring the scaling to reduce to maintain a targeted acceptance ratio of 25%.

Lastly, as starting values $\boldsymbol{X}^{\{0\}}$ I choose parameter estimates of the multinomial logit model estimated on the aggregate of all segments departing the same state.

4.6.1 Burn-in and sample length

For the segments departing CURRENT I use 150,000 burn-in iterations after which I will sample 100,000 pulls from the posterior. For the segments departing DELINQUENT 30-89 DAYS and DELINQUENT 89+ DAYS I use 75,000 burn-in iterations and sample 75,000 pulls from the posterior, as can be seen in Figure 4.5. These choices were made whilst investigating select traces of the sampled parameters on convergence. Additionally it was taken into account that the segments departing CURRENT have a lot more parameters and might be less stable, they therefore have a longer burn-in period and larger sample size. Lastly the practicality of the simulation were considered, such that exceedingly large number of samples were not possible.

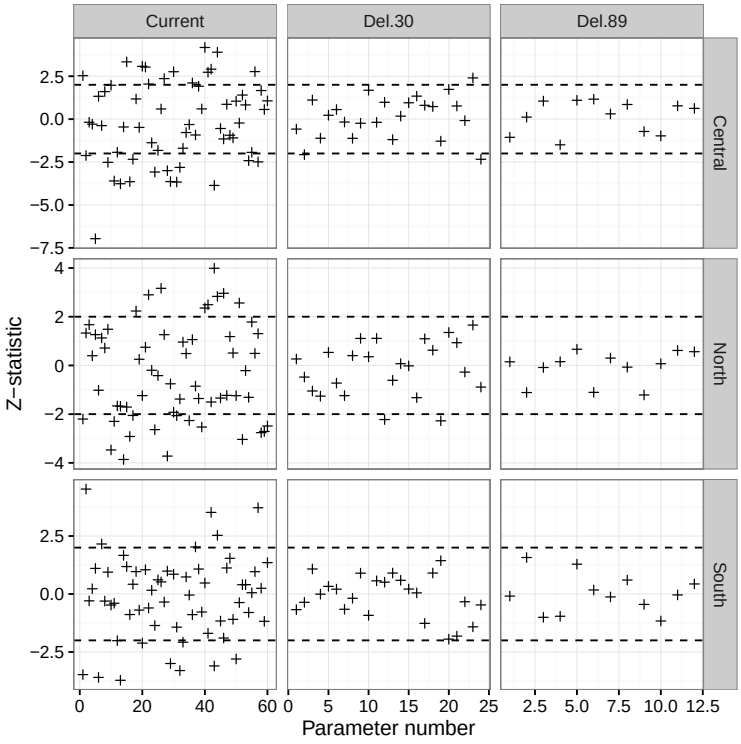
To determine whether the number of burn-in iterations and the number of samples generated from the adaptive Metropolis-Hastings algorithm were sufficient I make use of the work of Geweke (1992). Geweke (1992) proposes to test the convergence of MCMC samplers by analyzing the stability of the distribution of the sampled parameters.

Divide the generated sampled paths (post burn-in) into 2 subsets, A and B, n_A and n_B being the number of pulls present in each respective sub-set. Take A to be the first 20% of the generated sample, and B to be the last 50% of the sample. These percentages are in-line with what Geweke (1992) advocate. The important thing is that the subsets do not overlap and are sufficiently apart from each other to be considered independent. Under convergence of the algorithm it would be expected that the pulls in these two subsets are from the same distribution. To test this Geweke (1992) proposes the following test statistic:

$$z = \frac{(\bar{\theta}_{\mathrm{A}} - \bar{\theta}_{\mathrm{B}})}{\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}} \sim N(0,1) \tag{4.24}$$

Where $\bar{\theta}_{\mathrm{A}} = n_{\mathrm{A}}^{-1} \sum_{n \in A} \theta^{\{n\}}$ is the sub-set mean and σ_A^2 the sub-set variance[9]. This statistic is calculated for all parameter paths.

Figure 4.6: Z-stats sampled parameter values



Calculated Z-statistic for the Geweke (1992) MCMC convergence test. Each cross is a model parameter. Dashed lines indicate the 5% significance level.

The results of these z-statistics can be found in Figure 4.6. Each cross indicates a model parameter. The parameters are grouped by departing state and geographic segmentation. As can be seen is that the vast majority of parameters have converged in distribution. Although it should be noted that for the parameters in the model segments departing CURRENT some parameters are outside the 5% significance level. This indicates that for this departing state the parameter estimates would probably benefit from a longer burn-in or a longer post burn-in period. Due to pragmatic limitations (i.e. run time), the trade-off was made to maintain a 150,000 burn-in period and a 100,000 sample size. For the segments departing the non-performing states DELINQUENT 30-89 DAYS and DELINQUENT 89+ DAYS we see clear convergence of the sampled distribution, in-line with what can be seen in Figure 4.5.

[9]The variance is estimated taking into acount the inherent auto-correlation in the samples. This is done by estimating $\sigma_A^2 = n_{\mathrm{A}}^{-1} S_A^{(0)}$, where $S_A^{(0)}$ is the spectral density at 0. An AR model can be used to determine $S_A^{(0)}$, see Geweke (1992) for more details.

5 General setup and Model comparison measures

I estimate the models on mortgages active at the start of 2010 and their respective performance states at the end of 2010. With this I mean that the mortgages that are not in DEFAULT or PREPAID state at the end of December 2009 are considered active as of the start of 2010. The mortgage characteristics and the constructed macro economic variables, as outlined in Section 3.2, per the beginning of 2010 will then form the data-set of explanatory variables. A mortgage is considered to have transitioned to DEFAULT or PREPAID if the mortgage defaults or prepays at any point during 2010. A transition to the non-preforming states DEL. 30-89 DAYS and DEL. 89+ DAYS is noted as a mortgage that is 30-89 days or 89+ days behind on payments at the end of 2010. I then take the estimated model parameters and apply this to mortgages active at the start of 2011 and evaluate the forecasts of the performance state of each mortgage to the observed transition.

The mortgages active at the start of 2010 are divided, randomly, into two subsets. $\frac{2}{3}$ of the mortgages are used to estimate the models, $\frac{1}{3}$ of the mortgage are then used to calibrate the hyper-parameter g for Bayesian multinomial logit models. The value of g resulting in the best forecasts for this separate group of mortgages is then used.

As benchmark a naive approach to forecasting mortgage performance will be applied. The benchmark transition probabilities for all mortgages will be equal to the proportion of observed transitions in the estimation sample, $\hat{\boldsymbol{\pi}}_m = \boldsymbol{\pi}_0$, this is equivalent to a multinomial logit model with only intercepts and no other explanatory variables. Note that the implied decision rule for the benchmark model is then equal to randomly forecasting a transition taking place with probability $\boldsymbol{\pi}_0$ as all mortgages have the same set of transition probabilities[10]. This has a couple of advantages. First, in forecasting mortgage performance a simple approach is to use past performance, which is what the benchmark model does. Secondly, when looking at in-sample and out-of-sample hit-rate the benchmark model is what is tested against for significance, making for easy comparison.

I implement Bayes factors to compare in-sample evidence of the Bayesian multinomial logit model to the benchmark and standard (non-segmented) multinomial logit. Bayes factors take into account the increase in parametrization of the models and do not only look at increase in in-sample fit. Additionally, given the Bayesian nature of the research it makes for a natural comparison measure. Bayes factors are further expanded up in Section 5.1.

Additionally, to compare the models I make use of hit-rates for in and out-of-sample accuracy. The choice for using hit-rates, although their frequentist nature, was made because it is the most natural and easiest way of interpreting mortgage performance forecasts. At the end of the day one can say if a forecast for a mortgage was correct or not. Section 5.2 goes into more detail.

The McFadden R^2 will also be used as an in-sample fit measure when comparing the multinomial logit model, estimated via maximum likelihood, to the benchmark.

$$R^2 = 1 - \frac{\log(L_{\text{ML}})}{\log(L_{\text{BM}})} \tag{5.1}$$

This has as advantage that the McFadden R^2 measures the increase of log likelihood over that which can be achieved under the naive benchmark model of only including intercepts. A positive value would indicate increase in model fit. However it should be noted that the McFadden R^2 does not account for number of parameters and as such is just used as an initial indication of increase in model fit.

[10] The hit-rate for the benchmark can then be shown to be $h_{bm,j} = \pi_{0,j}\pi_{f,j} + (1-\pi_{0,j})(1-\pi_{f,j})$

5.1 Bayes factors

For the Bayesian multinomial logit model I make use of Bayes factors. These are calculated as the ratio of the evidence of the data under the considered model to that under the benchmark. This is also known as the ratio of marginal likelihoods of the models. To calculate these likelihoods the unknown parameters are integrated out.

$$\begin{aligned} \text{BF}_{A|B} &= \frac{p_A(y)}{p_B(y)} \\ p_A(y) &= \int_{\theta_A} p_A(y|\theta_A)p(\theta_A)d\theta_A \end{aligned} \tag{5.2}$$

This has the advantage that the evidence is not based on a single outcome (estimation) of the parameter value, but the entire prior parameter distribution is taken into account. Kass & Raftery (1995) provide a approximation to calculating Equation 5.2:

$$\log(\text{BF})_{A|B} \approx \log(p_A(y|\hat{\theta}_A)) - \log(p_B(y|\hat{\theta}_B)) - \frac{1}{2}(k_A - k_B)\log(M) \tag{5.3}$$

Where $\hat{\theta} = \frac{1}{G}\sum_{g=1}^{G}\theta^{\{g\}}$, the average posterior sampled θ in the adaptive Metropolis-Hastings algorithm of Section 4.6.

5.2 Binomial and multinomial hit-rate

To be able to compare the in and out-of-sample accuracy of the models I implement a hit-rate measure. I propose the use of the following decision rule to translate the estimated transition probabilities $\hat{\boldsymbol{\pi}}$ to actual forecast transitions $\hat{x}^i_{m,j}$ for each mortgage. Given some set of mortgages $m \in M^{i,\mathcal{L}}$ departing state i in segment $\mathcal{L}$ the binomial decision rule for each of the possible arriving states j is:

$$\hat{x}_{m,j} = \begin{cases} 1 & \text{if } \hat{\pi}_{m,j} \geq \pi^*_j \\ 0 & \text{if } \hat{\pi}_{m,j} < \pi^*_j \end{cases} \tag{5.4}$$

Where π^*_j is the $(1-\bar{\pi}_j)$th quantile of the forecast transition probabilities to state j of all mortgages m departing state i in segment $\mathcal{L}$. $\bar{\pi}_j$ is the average transition probability of the respective mortgages. For example, assume that the average transition probability $\bar{\pi}_{\text{default}} = 2\%$. The 2% of the mortgages with the highest probability of transitioning to DEFAULT are forecast to make this transition, the other 98% not. This decision rule has the advantage that the transition forecasts are consistent with $\bar{\boldsymbol{\pi}}$, the average probabilities. Additionally the decision rule takes into account the probability of each forecast transition.

$$w_{m,j} = \begin{cases} 1 & \text{if } \hat{x}_{m,j} = x_{m,j} \\ 0 & \text{else} \end{cases} \tag{5.5}$$

The binomial hit-rate is then $h_j = \frac{1}{M}\sum w_{m,j}$. To test if the hit-rate is significantly higher than that which can be expected under independence the following test statistic is calculated:

$$z = \frac{h_j - q_j}{\sqrt{q_j(1-q_j)/n}} \sim N(0,1) \tag{5.6}$$

with $q_j = \pi_{f,j}^2 + (1 - \pi_{f,j})^2$ the random hit-rate achieved under the sample probabilities $\boldsymbol{\pi}_f$[11]. See also Heij *et al.* (2004) and Pesaran and Timmermann (1992) for more comprehensive discussions on binomial hit-rate.

In addition to this binomial hit-rate per possible arriving state, I also implement a multinomial hit-rate measure to compare overall model accuracy, not just accuracy of forecasting a single arriving state. The hit rate then becomes $w_m = 1 \quad$ if $\hat{s}_m = s_m$ and zero otherwise. The difficulty in calculating this multinomial hit-rate measure is in defining a decision rule that reflects the forecasts made. To this end I propose a procedure that maps $\hat{\boldsymbol{\pi}}_m \mapsto \hat{s}_m$.

A common way to translate multinomial transition probabilities by setting $\hat{s}_m = \underset{j}{\text{argmax}}[\hat{\pi}_{m,j}]$, set the estimated value of the state to the that state for which the transition probability is highest. See for example Heij *et al.* (2004) and Gneiting & Raftery (2007). In the application of this research this would result in almost always estimating the state to which a mortgage transitions to be CURRENT for those mortgages departing CURRENT, as this ending state almost always has the highest estimated transition probability. The algorithm aims to assign transitions to mortgages such that:

1. $\frac{1}{M}\sum_m 1[\hat{s}_m = j] = \frac{1}{M}\sum_m \hat{\pi}_{m,j}$; the proportion of estimated transitions to state j is equal to the respective average forecast transition probability of state j.
2. $\hat{\pi}_{m,j} \geq \hat{\pi}_{n,j} \quad \forall\{m,n\} \quad$ where $\hat{s}_m = j$ and $\hat{s}_n \neq j$; all mortgages that get assigned a transition to state j have an equal or higher scaled transition probability of this transition than those mortgages that did not get assigned a transition to state j.

Define:

$$\hat{\boldsymbol{\Pi}} = \begin{bmatrix} \frac{\pi_{1,1}-\mu_1}{\sigma_1} & \cdots & \frac{\pi_{1,j}-\mu_j}{\sigma_j} & \cdots & \frac{\pi_{1,j}-\mu_J}{\sigma_J} \\ \vdots & & \vdots & & \vdots \\ \frac{\pi_{m,1}-\mu_1}{\sigma_1} & \cdots & \frac{\pi_{m,j}-\mu_j}{\sigma_j} & \cdots & \frac{\pi_{m,J}-\mu_J}{\sigma_J} \\ \vdots & & \vdots & & \vdots \\ \frac{\pi_{M,1}-\mu_1}{\sigma_1} & \cdots & \frac{\pi_{M,j}-\mu_j}{\sigma_j} & \cdots & \frac{\pi_{M,J}-\mu_J}{\sigma_J} \end{bmatrix} \tag{5.7}$$

a matrix with the scaled multinomial transition probability vectors. Where $\mu_j = \frac{1}{M}\sum_m \hat{\pi}_{m,j}$ and $\sigma_j^2 = \frac{1}{M}\sum_m(\hat{\pi}_{m,j} - \mu_j)^2$. Such that each element is scaled by the respective column mean and column standard deviation. Also define $\boldsymbol{c} = [c_1, \ldots, c_J] = \mathbf{0}$.

1. Find the largest element in $\hat{\boldsymbol{\Pi}}$, $\pi^{\max}_{m*,j*}$.
2. Set $s_{m*} = j*$ for the mortgage $m*$ and state $j*$ belonging to $\pi^{\max}_{m*,j*}$.
3. Set the $m*$th row of $\hat{\boldsymbol{\Pi}}$ to $-\inf$
4. Set $c_{j*} = c_{j*} + 1$
5. If $c_{j*} > \bar{\pi}_{j*}M$; set all elements in the $j*$th column of $\hat{\boldsymbol{\Pi}}$ to $-\inf$.
6. Repeat from step (1) M times such that all mortgages have an assigned transition s_m.

The above procedure will result in a state forecast $\hat{x}_m$ per mortgage, as well as maintain that the proportion of forecasts states are equal to the implied aggregated forecast probability. The hit-rate can then be calculated as with the binomial case by counting the "hits", defined as: $w_m = 1[\hat{s}_m = s_m]$. It can also be tested for significance via a generalization of Equation (5.8):

[11] $\pi_{f,j} = n_j/M$ defined as the observed proportion of transitions to j in the forecast sample

$$z = \frac{h - q}{\sqrt{q(1-q)/n}} \sim N(0,1) \tag{5.8}$$

Where $h = \frac{1}{M}\sum w_m$ is the multinomial hit-rate of the model and q is the hit-rate that would be expected if the observed sample probabilities were used to randomly assign transitions: $q = \sum_j \pi_{f,j}^2$.

6 Results

To evaluate the results I will first discuss the multinomial logit model estimated over all aggregated segments per departing state in comparison to the benchmark. This will allow for analysing the significance and interpretation of the covariates discussed in Section 3.2. After this I continue to discuss the in-sample fit of the multinomial logit model. I continue then to discuss the in-sample fit of the Bayesian multinomial logit model in Section 6.2. Although I will mainly concentrate on the geographic segmentation scheme, I will also touch upon the other 2 segmentation methods (by vintage and K-means). Results of less significance to the main narrative of this thesis will be presented in the Appendix A.1.

For clarity, in the result tables I maintain the following designation of models:

BM: Benchmark model. This always refers to a simple multinomial logit model per departing state with only intercepts and no other explanatory variables. Note that this is equivalent to using the observed ratio of transitions as estimates for the transition probabilities for all mortgages. As such the in-sample and out-of-sample hit-rates are tested for significance against this naive benchmark.

ML: Multinomial logit. This consists of one model per departing state and includes explanatory variables. Significance of in-sample fit and out-of-sample accuracy will be compared to the benchmark model to deduce if the explanatory variables add significant fit to the model. For a comparison of the multinomial logit model, without segmentation, to the Bayesian multinomial logit models, with segmentation, I also estimate the model via the use of the discussed sampling techniques. For this case I however set the prior distribution as un-informative with $g = 10.000$. Note that although this model is not estimated on a segment level, the parameter estimates are applied on the segments separately to allow for comparison with the Bayesian multinomial logit model more easily.

BML: Bayesian multinomial logit. This is the model that has been expanded with segmentation. It is formulated in a Bayesian manner with prior distributions applied to the parameter rows as discussed in Section 4.5 and estimated via the adaptive Metropolis-Hastings algorithm of Section 4.6. The Bayesian multinomial model, which has segmentation, is compared to the standard multinomial model, which does not have segmentation, to determine the added benefit of segmenting the mortgage pool. A comparison is also made with the benchmark model to determine if the explanatory variables still add significance over just the use of intercepts.

6.1 Multinomial logit estimates and in-sample accuracy

Table 7: ML estimates for transitions departing CURRENT

	Del. 30-89 days	Del. 89+ days	Default	Prepaid
(intercept)	$1,6795$	$-3,788^{**}$	$-4,6411^{*}$	$-4,1466^{***}$
	(1,1802)	(1,1882)	(2,3997)	(0,4421)
Int. rate diff.	$0,115$	$0,9306^{***}$	$0,5161$	$1,1004^{***}$
	(0,1473)	(0,227)	(0,389)	(0,069)
County unemp.	$-0,0463$	$-0,1133^{**}$	$-0,2921^{***}$	$-0,0968^{***}$
	(0,037)	(0,0367)	(0,0793)	(0,0127)
Estimated LTV	$0,0182^{***}$	$0,0345^{***}$	$0,05^{***}$	$-0,0097^{***}$
	(0,0026)	(0,0026)	(0,006)	(0,0009)
Ever del. TRUE	$1,4368^{***}$	$0,7987^{***}$	$1,1482^{***}$	$-0,4688^{***}$
	(0,1506)	(0,1686)	(0,3289)	(0,1012)
Time since prof	$-0,0035$	$-0,0222$	$-0,0491$	$-0,0627^{***}$
	(0,0126)	(0,0172)	(0,0392)	(0,007)
Credit score	$-0,0106^{***}$	$-0,0055^{***}$	$-0,0021$	$0,0043^{***}$
	(0,0013)	(0,0013)	(0,0027)	(0,0005)
DTI	$0,0166^{**}$	$0,0263^{***}$	$0,0275^{*}$	$-0,0016$
	(0,0056)	(0,0056)	(0,0113)	(0,0017)
Jumbo TRUE	$-0,1072$	$0,0169$	$-0,4278$	$0,6942^{***}$
	(0,1507)	(0,1479)	(0,3137)	(0,0482)
Log(loan age + 1)	$-0,0443^{*}$	$0,1375^{*}$	$-0,2722$	$-0,0871^{***}$
	(0,0914)	(0,1125)	(0,1959)	(0,0253)
Occ. stat O	$0,7122$	$0,6619$	$-0,3789$	$0,7738^{*}$
	(0,3233)	(0,2929)	(0,4697)	(0,0906)
Occ. stat S	$-0,37^{*}$	$-0,2033$	$0,3232$	$0,2844^{***}$
	(0,6628)	(0,5409)	(0,6528)	(0,1507)
Loan purpose N	$-0,3646^{**}$	$-0,1454^{**}$	$-0,2188^{*}$	$0,2394^{***}$
	(0,1781)	(0,178)	(0,3642)	(0,0512)
Loan purpose P	$-0,602$	$-0,5158^{***}$	$-0,7148^{**}$	$0,3906^{***}$
	(0,1789)	(0,1635)	(0,3218)	(0,057)
Num. borrowers 2	$-0,0891$	$-0,5213^{***}$	$-0,7909^{**}$	$0,2136^{***}$
	(0,1344)	(0,1328)	(0,2757)	(0,0447)

m = 15.214 mortgages; (***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels.
Conditioned on CURRENT as arriving state
LR-stat: $-9404,9^{***}$

Table 7 above shows the parameter estimates for the log-odds ($\log(\frac{\pi_j}{\pi_1})$) for mortgages departing from state CURRENT. A positive parameter value indicates that an increase in the respective covariate increases the odds of transitioning to the indicated ending state **over** a transition to CURRENT (as the CURRENT ending state is being conditioned upon). For example, a higher *Estimated LTV* increases the odds for a mortgage to transition to DEL. 30-89 DAYS, DEL. 89+ and DEFAULT the coming year over transitioning to CURRENT. Whereas *Credit score* decreases the odds of transitioning to one of the non-performing states, but increases the odds of a mortgage being PREPAID over being CURRENT.

When looking at the drivers behind a transition to PREPAID we can see a couple of intuitive outcomes. Namely, we notice that *interest rate diff.* and *credit score* effect the odds of prepayment in a significant positive manner. This is to be expected as a positive interest rate differential gives a financial incentive to the borrower to re-finance their mortgage, triggering in effect a prepayment event. Likewise, a higher *credit score* indicates that the borrower is more likely to be accepted for a refinance mortgage and is thus more likely to transition to PREPAID than to stay CURRENT compared to a borrower with a lower *credit score*.

In general we can see that covariates that indicate that the borrower is in an economically better position, *Occupancy status: Seconds home*, *Num. borrowers 2* and *Loan purpose* (either Purchase or No cash-out refinance), increase the odds of prepayment.

Conversely, *County unemp.* and *Log(loan age + 1)* have a negative effect on the odds to prepay. Note that *County unemp.* is a proxy for the economic well being of the respective county, as such a higher unemployment rate could indicate less economic room or opportunity for a borrower to refinance. The negative effect of *Log(loan age + 1)* on the odds to prepay could be a result of the fact that older mortgages will have repaid, via monthly redemption payments, more of the notional than younger mortgages. As such the financial incentive to refinance is lower. One can also consider the fact that if a mortgage has had a financial incentive to refinance, but has not done so, there might be other circumstances preventing the borrower from doing so. This is reflected by the significant negative impact of *Time since prof* on the odds to prepay over remaining CURRENT.

Of interest is the result that higher *interest rate diff.* also increases the odds of transitioning to the non-performing state DEL. 89+ DAYS. A possible explanation of this can be that ineligibility to profit from a lower interest rate environment can signal that the borrower is not dot doing so well financially. The significant negative relation between *county unemp.* and the odds of transitioning to DEL. 89+ DAYS and DEFAULT is especially interesting. One would expect increasing unemployment rates to increase the odds of transitioning to these non-performing states. It should be noted however that the parameters were estimated cross-sectional, as such rising unemployment in a unique county is not captured. What could be happening in this case is that counties with high unemployment are already subject to stricter underwriting criteria as these areas are known to be economically weaker. This is still in-line with the negative effect of *county unemp.* on prepayment, as these economically weaker counties, although subject to stricter underwriting, are not in a better position to prepay.

Table 8: ML estimates for transitions departing DELINQUENT 30-89 DAYS

	Del. 30-89 days	**Del. 89+ days**	**Default**	**Prepaid**
(intercept)	$2,8362$	$0,5507$	$-6,2411^{*}$	$-7,2713$
	(1,7487)	(1,6191)	(2,8946)	(4,5418)
Int. rate diff.	$-0,0679$	$0,369^{*}$	$0,5288$	$1,0988^{*}$
	(0,2142)	(0,2054)	(0,3527)	(0,5569)
County unemp.	$0,0578$	$-0,1148^{*}$	$-0,0752$	$0,0656$
	(0,0698)	(0,0659)	(0,1069)	(0,1621)
Estimated LTV	$-0,0029$	$0,0066$	$0,0207^{**}$	$-0,0424^{***}$
	(0,0049)	(0,0046)	(0,0076)	(0,0124)
Credit score	$-0,0062^{**}$	$-0,0014$	$0,004$	$0,0087$
	(0,0023)	(0,0021)	(0,0038)	(0,006)
Jumbo TRUE	$0,5234^{*}$	$0,2363$	$-0,3011$	$0,1543$
	(0,2597)	(0,2369)	(0,432)	(0,6323)

m = 511 mortgages; (***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels.
Conditioned on CURRENT as arriving state
LR-stat: $-629,17^{***}$

Tables 8 and 9 show the parameter estimates of the log-odds for the states departing DELINQUENT 30-89 DAYS and DELINQUENT 89+ DAYS. First we note that reduction in covariates that entered these models. This is caused by two elements. First there are far fewer mortgages departing these states as there are mortgages departing the state CURRENT. Second, as mortgage loans become non-performing the drivers that govern the actions of the borrowers under normal circumstances disappeared. To give an example of

the latter, let us notice that the effect of *interest rate diff.* has weakened for the mortgages departing state DEL. 30-89 DAYS and become entirely non-significant for those mortgages departing state DEL. 89+ DAYS. As a borrower becomes more behind on their monthly payments, it is conceivable it would become more difficult to meet refinancing conditions, even if it is financially optimal to do so. At the least it would become more expensive, as a higher rate might apply due to the bad payment performance of the borrower.

Table 9: ML estimates for transitions departing DELINQUENT 89+ DAYS

	Del. 30-89 days	**Del. 89+ days**	**Default**	**Prepaid**
(intercept)	$7,5348^{**}$	$3,5799^{**}$	$-0,5878$	$7,8885$
	(2,579)	(1,3533)	(1,4399)	(6,5304)
Estimated LTV	$-0,0119^{*}$	$-0,0014$	$0,0081^{*}$	$-0,0721^{**}$
	(0,0066)	(0,0034)	(0,0036)	(0,0222)
Credit score	$-0,0118^{**}$	$-0,0047^{*}$	$-0,0005$	$-0,0084$
	(0,0037)	(0,0019)	(0,002)	(0,0092)

m = 703 mortgages; (***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels.
Conditioned on CURRENT as arriving state
LR-stat: $-875,25^{***}$

Of main importance is that only *Estimated LTV* and *Credit score* remain of some significance to the odds of the transitions. Where *Estimated LTV* is mainly of explanatory value to the odds of prepayment and default. Intuitively, if a loan is non-performing, but the property is still worth significantly more than the outstanding notional, a mortgage provider might be willing to refinance the loan or work out an adjusted payment plan, most likely under stricter conditions, due to the risk mitigating equity in the home. Likewise, if a mortgage has negative equity, and is non-performing, it becomes more probable that the borrower just "walks away" from their mortgage obligation. Similarly, negative equity and non-performance might trigger the mortgage provider to demand immediate repayment, where otherwise there might be room for an negotiated payment plan. This would also trigger a default event.

Lastly of note is that *credit score* is no longer of significance to the odds of a transition to the DEFAULT state once a mortgage becomes non-performing. This could be due to the fact that CREDIT SCORE is a snap-shot of the credit worthiness of the borrower. Once the mortgage is non-performing this snap-shot is no longer representative of the credit worthiness of the borrower. For example, a borrower might not be able to make payments due to unemployment or unexpected other expenses (e.g. injury).

Figure 6.1: ML transition estimates

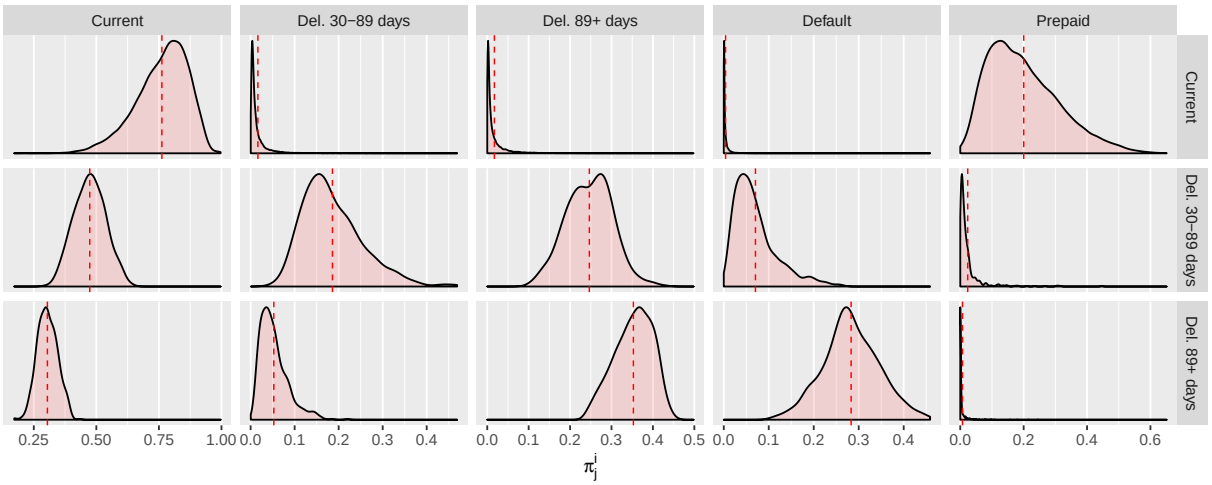


Figure shows the distribution of fitted transition probabilities, $p(\hat{\pi}^i_{ML,j})$, of transitioning from a departing state i (rows) to an ending state j (columns). Dashed line indicates the benchmark transition estimates π_0. Note that the y-axis has been scaled for all plots to increase readability, as such the relative heights are not comparable.

Figure 6.1 above summarizes the in-sample dynamics of the mortgages. It shows the distribution of fitted in-sample transition probabilities per departing/ arriving state combination. What can be see is that departing from the performing state of CURRENT mortgages have a broad range of probabilities with respect to arriving in CURRENT and PREPAID. Where as an arrival in the non-performing states are far more isolated. This changes drastically once the mortgages start in one of the non-performing states DELINQUENT 30-89 DAYS and DELINQUENT 89+ DAYS. This, together with the parameter estimates, confirms the assumption of the Markovian structure in mortgage performance. With this I mean that the distribution of the multinomial probabilities is dependent on the departing state, and differs significantly between the 3 different departing states. The dashed line indicates the observed ratio of transitions. This is equivalent to the benchmark model, which assumes the constant transition probability vector for all mortgages departing the same state. It can be seen that although the fitted ML transition probabilities are concentrated around the average observed transitions, there can be a large degree of variation, setting the stage for increasing individual mortgage transition forecast accuracy.

In-sample accuracy

Table 10 below summarizes the in-sample performance of the multinomial logit model. Although the performance is also show on a segment level, the model itself was estimated on the aggregate mortgages departing each of the three states CURRENT, DEL. 30-89 DAYS and DEL. 89+ DAYS. The stated in-sample performance are derived by applying the parameter estimates of the aggregate data of each departing state to each segment. The multinomial logit models is tested for significance against the benchmark, which assumes the ratio of observed transitions as the probability vector for each mortgage. Such that each mortgage is assumed to have the same probability vector. Which is synonymous to a multinomial logit model with only intercepts.

First it can be noted that the addition of the explanatory variables to the multinomial logit improves in-sample fit. Not that all of the aggregated multinomial hit-rate scores are significantly higher than that which would be expected under the benchmark model. For the mortgages departing CURRENT and DEL.

30-89 DAYS the multinomial logit model manages an in-sample hit-rate of roughly 7% points higher than the benchmark.

Table 10: Multinomial logit in-sample accuracy

	Binomial hit-rate (transition to):					Multinomial hit-rate		
Departing State	Current	Del. 30-89	Del. 89+	Default	Prepaid	ML	BM	R^2
Current	69,6%***	97,2%***	97%***	99,3%	74,8%***	68,9%***	62,1%	0,103
Central	69,2%***	97,7%	97,2%	99,2%	73,8%***	68,2%***	61,6%	0,111
North	73,1%***	96,9%	96,2%	99,7%	77,3%***	70,8%***	64,2%	0,09
South	69,4%***	96,9%***	96,9%**	99,2%	74,8%***	68,8%***	62,2%	0,099
Del. 30-89 days	54,2%**	72,2%	65,9%*	88,6%	96,1%	39,5%***	32,5%	0,043
Central	51,7%	71,4%	64,6%	85,7%	97,3%	34%	32,2%	0,034
North	62,5%*	78,1%*	81,3%*	78,1%	93,8%	43,8%**	30,4%	0,059
South	55,7%**	71,7%	64,8%	90,4%	95,2%	39,8%***	32,9%	0,046
Del. 89+ days	60,7%**	90,9%	59,9%***	63%**	98,6%	33,3%**	30,0%	0,025
Central	60,5%	93%	64,7%***	61,9%	99,1%	41,9%***	30,7%	0,044
North	72,4%	86,2%	55,2%	51,7%	96,6%	34,5%	27,8%	-0,004
South	60,3%*	90,4%	58,6%**	62,3%	98,5%	32,2%	29,8%	0,019

(***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels. Binomial hit-rate is only shown for the multinomial logit model.

If we take a look at the binomial hit-rates, measuring how well the model can predict the respective binary event, it can be seen that the multinomial logit model is mainly exceeding the benchmark when it comes to the CURRENT and prepaid arriving states and then primarily when CURRENT is the departing state. It should be noted that these events occur relatively the most often, therefore there are more observations to allow for a better estimation of the parameters. Especially transitions to the DEFAULT state are not fitted significantly better than the benchmark. Note however that whilst a hit-rate in excess of 99% looks impressive, it is not significantly better than the benchmark hit-rate. For events with extreme probabilities, close to 0% or 100%, these high hit-rates are to be expected.

Overall it can be stated that the multinomial logit model manages to improve on the naive benchmark. This is also reflected in the McFadden R^2 statistic, which denotes a positive value for all departing states. Although this improvement in fit is mainly caused by mortgages departing from CURRENT and arriving in the performing states CURRENT and PREPAID. Specifically, the multinomial logit model fits well for those transitions for which there are plenty of observations, but struggles with events that are less likely to occur.

6.2 Bayesian Multinomial Logit

To assess the performance of the Bayesian multinomial logit model I will first look at the in-sample performance. Note that the hyper parameter governing prior variance, g, was first to be determined. This was done by splitting off $\frac{1}{3}$ of the estimation sample and using this for in-sample forecasting. Based on the in-sample forecast accuracy the hyper parameter g was chosen as the value that maximized in-sample multinomial hit-rate. In general it was found that segments with a large number of observations would have a higher optimal g. This implies that the prior is less informative in this situation and the parameters estimates are shrunk less toward that of the maximum likelihood estimates of the multinomial logit model on the aggregated mortgages of all segments departing from the same state. This is intuitive as these segments

contain enough observations to allow for accurate estimation of the parameters, without needing information from the aggregated mortgage pool with the same starting state.

Segments with fewer observations, those segments departing from non-performing states DEL. 30-89 DAYS and DEL. 89+ DAYS, were found to prefer a lower value of g. This indicates the application of a more informative prior. In a sense these segments are allowing more information from the aggregated mortgages to enter the model. The parameter estimates are thereby shrunk toward the estimates of the multinomial logit model. Note however that, with the exception of the SOUTH segment for departing state DEL. 89+ DAYS, the specification of g in the optimal setting was never below 1. This indicates that the prior variance was never scaled such that the prior completely dominates the posterior distribution.

Table 11: Bayesian multinomial logit in-sample accuracy

	Binomial hit-rate (transition to):					Multinomial	$2\text{Log}(\text{BF})_{\text{BML}\mid *}$		
Departing State	Current	Del. 30-89	Del. 89+	Default	Prepaid	hit-rate	$*\mid$BM	$*\mid$ML	
Current	$70,0\%^{***}$	$97,2\%^{***}$	$97,0\%^{***}$	$99,3\%$	$74,8\%^{***}$	$69,0\%^{***}$	844,1	-485,7	
Central	$70,1\%^{***}$	$98\%^{***}$	$97,6\%^{***}$	$99,2\%$	$74,0\%^{***}$	$69,2\%^{***}$	543		g=10
North	$73,8\%^{***}$	$97,3\%$	$95,7\%$	$99,9\%^{***}$	$78,2\%^{***}$	$72,2\%^{***}$	-282,4		g=100
South	$69,4\%^{***}$	$96,7\%$	$96,8\%$	$99,3\%$	$74,9\%^{***}$	$68,5\%^{***}$	825,5		g=10
Del. 30-89 days	$56,0\%^{***}$	$72,2\%^{*}$	$66,3\%^{**}$	$87,9\%$	$95,7\%$	$36,8\%^{**}$	19,3	3,1	
Central	$53,1\%$	$71,4\%$	$63,9\%$	$85,0\%$	$97,3\%$	$34,0\%$	31,1		g=1
North	$71,9\%^{***}$	$75,0\%$	$81,3\%^{**}$	$75,0\%$	$93,8\%$	$31,3\%$	35,4		g=2
South	$55,7\%^{**}$	$72,3\%$	$66,0\%$	$90,4\%^{**}$	$95,2\%$	$38,6\%^{***}$	41,8		g=1
Del. 89+ days	$61,5\%^{**}$	$91,5\%$	$59,7\%^{***}$	$62,6\%^{**}$	$98,7\%$	$33,3\%^{**}$	207,5	70,4	
Central	$61,4\%^{***}$	$95,3\%^{***}$	$62,8\%$	$62,8\%$	$99,1\%^{***}$	$34,0\%$	76,2		g=2
North	$86,2\%^{***}$	$86,2\%^{**}$	$51,7\%$	$51,7\%$	$96,6\%$	$37,9\%$	67,1		g=2
South	$59,9\%^{***}$	$90,0\%^{***}$	$58,8\%$	$63,2\%$	$98,7\%^{***}$	$32,7\%$	102,6		g=0,1

(***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels.

When considering the in-sample hit-rate we find largely the same results as were found for the multinomial logit model. Given that both models use the same set of explanatory variables this is to be expected. We once again see that fitting of the transitions departing the performing state CURRENT to the arriving states CURRENT and PREPAID are modeled relatively well in comparison to the benchmark. A general improvement to the in-sample hit-rate for the Bayesian multinomial logit model over the multinomial logit model can be seen, but this should be expected as the number of estimated parameters has tripled.

To examine if the increase in parameters is justified by the increase in model fit, I look at the Bayes Factors. Firstly in comparison to the naive benchmark and secondly in comparison the multinomial model. As mentioned in Section 5, the log Bayes factors were estimated following the work of Kass & Raftery (1995). Kass & Raftery (1995) provide general guidelines to interpret these factors in the following manner:

Table 12: Interpretation of Bayes Factors

$2\log(\text{BF}_{A\mid B})$	$\text{BF}_{A\mid B}$	Evidence against M_B
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very Strong

To compare to the benchmark, the bayesian multinomial logit model was specified with only intercepts and run with an uninformative prior achieved by setting $g = 10.000$. Note that this differs slightly from the setup in the multinomial logit model, where the benchmark, as well as the multinomial logit model, were not estimated on a segment level but on the aggregated data per departing state. Due to the fact that the Bayesian multinomial logit does model each segment separately, it is only fair to use a benchmark that also does so[12].

In Table 11 it can be seen that the log Bayes Factors versus the benchmark, $2\text{Log(BF)}_{\text{BML|BM}}$, indicate strong evidence for the Bayesian multinomial logit model for most segments. This shows that the inclusion of the explanatory variables in these segments is to be preferred over the benchmark where no extra covariates are taken into account. Segments CENTRAL and SOUTH departing from CURRENT show strong evidence in favor of the more parameterized model. These segments also enjoy ample observations to be able to estimate the parameters. However, the NORTH segment in this same group departing from CURRENT shows no evidence in favor of the more complex model, instead it favors the parsimony of the benchmark. Note however that the Bayesian multinomial logit model for this departing state includes 60 parameters, where as the benchmark has 4, and this segment contains a lot fewer observations than the other two segments. See also Table 13 below. The evidence of model improvement in this segment is not enough to off-set the increased complexity of the model.

Table 13: Model parameterization

Departing State	M	k_{BM}	k_{ML}	k_{BML}
Current	15214	12	60	180
Central	5707	4		60
North	959	4		60
South	8548	4		60
Del. 30-89 days	511	12	24	72
Central	147	4		24
North	32	4		24
South	332	4		24
Del. 89+ days	703	12	12	36
Central	215	4		12
North	29	4		12
South	459	4		12

$k_{\dots}$ indicates the number of parameters in each of the three models.
BM and BML are estimated on a per segment basis. ML is only estimated per departing state.

A more interesting comparison is that versus the multinomial logit model. The main difference between BML versus ML is that BML imposes the segmentation structure where ML does not. Table 11 shows the log Bayes factors versus the multinomial logit model, $2\text{Log(BF)}_{\text{BML|ML}}$. The Bayes factors for the ML model were calculating by formulating the multinomial logit model as a Bayesian multinomial logit model with an uninformative prior, $g = 10.000$, estimated over the aggregated mortgages per departing state.

Where the BML model showed very strong evidence over the benchmark model for mortgage departing

[12]It should be noted for completeness that a benchmark defined as a Bayesian multinomial logit model with an uninformative prior using only intercepts applied to the aggregate mortgages per departing state does not result in a different conclusion than stated here with a benchmark per segment. The main difference being less parameters for such an aggregated benchmark, which could in theory result in no strong evidence in favor of the Bayesian multinomial logit model with explanatory variables, but does not do so in practice.

CURRENT, there is no evidence in favor of BML over the ML model. In fact, the increase in parameters for the BML model does not result in sufficient in-sample evidence and the more parsimonious ML model should be preferred. A more practical interpretation of this result is that for performing mortgages, there is no strong evidence for heterogeneity in the sample, at least not at a geographical level. For the mortgage departing state DEL. 30-89 DAYS it can be seen that there is positive evidence of the BML model over the ML model. Note that the evidence is not as strong as versus the benchmark model. Yet based on this data it seems that there is evidence supporting segmentation of the mortgages for this departing state. Indicating that the cost of having to estimate extra parameters, is offset by better capturing the heterogeneity in this group. Lastly it can be seen that for mortgage departing the state DEL. 89+ DAYS there is very strong evidence in favor of the BML model over the ML model. Once again, this implies that the addition of the segmentation structure, together with information from the aggregated group, allowed for capturing the heterogeneity present for these mortgages.

Vintage and K-means segmentation

In Table 14 the results for in-sample fit for segmenting by vintage can be found. We also see here that there is no evidence over the ML model to prefer segmentation by vintage for the departing state CURRENT. Just as with the geographical segmentation we notice that in-sample the vintage segmentation shows evidence for a better fit when considering the non-performing departing states DELINQUENCY 30-89 DAYS and DELINQUENCY 89+ DAYS.

Table 14: Bayesian multinomial logit in-sample accuracy, segmented by Vintage

	Binomial hit-rate (transition to):					Multinomial	2Log(BF)		
Departing State	Current	Del. 30-89	Del. 89+	Default	Prepaid	hit-rate	BML|BM	BML|ML	
Current	69,7%	97,2%	97%	99,3%	74,7%	69%***	347,0	-1195,0	
Crisis	68,2%***	95,5%	92,6%	98,4%	78,9%***	66,8%***	-122,6		g=100
Early	69,6%***	97,6%	98,8%***	99,9%**	72,2%***	69,2%***	-199,0		g=10
Post-crisis	70,4%***	98,2%***	98,2%***	99,5%***	73,1%***	69,8%***	495,1		g=5
Pre-crisis	69,4%***	96,4%	96,3%	99,2%	75,9%***	68,6%***	572,5		g=10
Del. 30-89 days	56,9%	72,4%	67,3%	90,4%	96,5%	39,7%***	76,8	13,8	
Crisis	63,6%***	77,3%**	63,6%	88,2%	99,1%**	37,3%	-20,8		g=10
Early	75%*	87,5%	75%	87,5%	87,5%	62,5%**	73,7		g=1
Post-crisis	56,9%*	79,6%***	68,6%*	88,3%	92,7%	40,9%**	38,6		g=1
Pre-crisis	53,5%	66%	68%**	92,6%***	97,7%**	39,5%***	139,4		g=0,01
Del. 89+ days	60,7%	90,6%	58%	63,3%	98,9%	33,9%**	221,8	8,1	
Crisis	58,6%***	88,9%***	56,1%	59,6%	99,5%***	29,3%	35,0		g=10
Early	58,3%	66,7%	33,3%	66,7%	91,7%	8,3%	81,7		g=1
Post-crisis	61,6%***	92,1%***	59,6%	62,9%	98,7%**	35,8%	71,2		g=1
Pre-crisis	61,7%***	91,8%***	59,4%	65,5%	98,8%***	36,5%*	94,6		g=1

(***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels.

For vintage segmentation is found that the CURRENT mortgages behave fairly similar to each other, irrespective of vintage year. This underlines the conclusion that performing mortgages are broadly driven by the same factors and react in the same manner to these factors. It was found that the explanatory variables added significant in-sample fit for this departing state, as such they do allow to be modeled better

than a naive benchmark. Yet the addition of segmentation has not yielded better results, and the relative parsimonious multinomial model is to be preferred. It can also be seen in Table 14, when looking at the log Bayes Factors, that for non-performing loans there is some in-sample evidence for differences in behaviour of mortgages from different origination years. This could be due, in part, by the fact that mortgages from the CRISIS and PRE-CRISIS segments have a relatively high loan-to-value ratio (due to housing prices decline during, and following, the crisis). This means that when these loans start to get behind on payments there is less room for an adjusted payment plan than a mortgage with a lot of positive equity. More lenient underwriting requirements also cause mortgages to be less resilient to economic set backs.

Both the geographic and vintage segmentation noted evidence of increased fit at the non-performing states. Motivating the separate modeling of these different segmentations. This is in contrast to the K-means segmentation, for which the in-sample results can be found in Table 15.

Table 15: Bayesian multinomial logit in-sample accuracy, segmented by Vintage

	Binomial hit-rate (transition to):					Multinomial	2Log(BF)		
Departing State	Current	Del. 30-89	Del. 89+	Default	Prepaid	hit-rate	BML|BM	BML|ML	
Current	70%	97,2%	97%	99,3%	75,1%	$69,3\%^{***}$	-405,8	-1483,6	
Normal	$66,7\%^{***}$	95,5%	94%	98,4%	$75,2\%^{***}$	$64,5\%^{***}$	261,0		g=100
Low Risk	$71,9\%^{***}$	$98,8\%^{***}$	$99,8\%^{***}$	$99,7\%^{***}$	$73,3\%^{***}$	$71,9\%^{***}$	-196,0		g=10
Mature	$67,6\%^{***}$	$98,2\%^{***}$	$99\%^{***}$	$99,9\%^{***}$	$70\%^{***}$	$67,5\%^{***}$	-202,2		g=5
High Risk	$76,3\%^{***}$	95,3%	93,3%	98,9%	$87,4\%^{***}$	$75,9\%^{***}$	49,7		g=10
Del. 30-89 days	56,2%	73,4%	67,7%	88,6%	96,3%	$40,1\%^{***}$	-101,6	-165,0	
Normal	$59,7\%^{***}$	73,1%	63,4%	87,8%	96,2%	$37,8\%^{**}$	-4,0		g=10
Low Risk	$70,8\%^{**}$	$83,3\%^{*}$	75%	$95,8\%^{*}$	95,8%	$58,3\%^{***}$	16,9		g=1
Mature	48,3%	73%	$71,9\%^{**}$	$97,8\%^{***}$	92,1%	$43,8\%^{**}$	-4,6		g=1
High Risk	53,1%	72,5%	$70,6\%^{**}$	83,8%	$98,8\%^{**}$	$38,8\%^{**}$	24,7		g=0,01
Del. 89+ days	61,2%	90,9%	59,6%	62,4%	98,4%	$35,6\%^{***}$	200,2	1,8	
Normal	$63,1\%^{***}$	$92,4\%^{***}$	57,3%	64%	$98,8\%^{***}$	$37,2\%^{**}$	50,0		g=10
Low Risk	$64,3\%^{*}$	$92,9\%^{***}$	60,7%	71,4%	92,9%	35,7%	81,3		g=1
Mature	$62,5\%^{**}$	$79,2\%^{*}$	64,6%	79,2%	91,7%	39,6%	66,2		g=1
High Risk	$58,3\%^{***}$	$90,8\%^{***}$	61,5%	56,9%	$99,6\%^{***}$	32,9%	63,0		g=1

(***): $\leq 1\%$, (**): $\leq 5\%$, (*): $\leq 10\%$ - significance levels.
Binomial hit-rate is only shown for the multinomial logit model.

Here it can be seen via the log Bayes Factors that the evidence of increased fit for the non-performing loans as deteriorated. Where for geographic and vintage segmentation the DELINQUENCY 30-89 DAYS showed evidence to be better modeled separately, we do not see any evidence for this under the K-means segmentation scheme. Even though the DELINQUENCY 89+ DAYS shows a positive log Bayes Factor, it is barely worth a mention. It seems that the K-means algorithm was able to find different clusters in the mortgage pool, but mortgages in these clusters do not react significantly differently to the explanatory factors than mortgages in other clusters. Where for geographic and vintage it was found that there was a basis, intuitive and statistical, to expect this difference in relation to the explanatory variables, for K-means this is not the case. In short, the clusters that were found in Section 4.4 might denote different risk groups, but the mortgages are still best modeled together. I must admit that this result is a bit disappointing, as the data-driven clusters that were found allowed for such an elegant interpretation.

On a general note, for both vintage and K-means segmentation we did see that the in-sample optimized g was less informative for larger segments and more informative (smaller g) for smaller segments.

6.3 Forecast accuracy

For the forecasts the parameters estimated over the 2010 data were applied to mortgages active at the start of 2011, and using the decision rules outlined in Section 5 a forecast state transition was made per mortgage. The forecasts transitions were then compared to what was actually observed and the hit-rates calculated. The results can be seen in Table 16, where the significance of the hit-rates is tested against the benchmark implied hit-rate.

Broadly we can see that both the ML and BML model manage to significantly outperform the benchmark in the case of forecasting mortgages departing CURRENT. As was already noted in when discussing in-sample performance, the performing state CURRENT contains enough observations to accurately estimate the explanatory variables and improve on forecast accuracy. When considering comparing the BML and ML model for this departing state it can be seen that the BML model performs slightly better than the ML model, although we found no evidence to prefer this model in-sample. Note however that the in-sample evidence takes into account that added parameterization of this model. Additionally the increase in forecast accuracy is only slight.

Table 16: 2011 transition forecast accuracy

ML forecast performance

Departing State	Binomial hit-rate (transition to): Current	Del. 30-89	Del. 89+	Default	Prepaid	Multinomial hit-rate	
Current	$72,4\%^{***}$	$97,5\%^{***}$	$97,6\%^{***}$	$99,4\%^{***}$	$76,8\%^{***}$	$71,8\%^{***}$	
Central	$71,5\%^{***}$	$97,7\%^{***}$	$97,5\%^{***}$	$99,3\%^{*}$	**76,0%*****	$71,0\%^{***}$	
North	$76,6\%^{***}$	$97,2\%$	**97,5%****	$99,4\%^{***}$	$81,6\%^{***}$	$76,1\%^{***}$	
South	$72,6\%^{***}$	**97,4%*****	**97,7%*****	**99,5%*****	$76,9\%^{***}$	$72,0\%^{***}$	
Del. 30-89 days	**50,3%**	**65,3%**	$68,1\%^{***}$	$87,5\%$	$93,9\%$	$33,3\%$	
Central	$50,3\%$	$67,8\%$	**63,7%**	$88,3\%$	$94,7\%$	$34,5\%$	
North	$39,0\%$	**68,3%**	$63,4\%$	$90,2\%$	$97,6\%$	**36,6%**	
South	**49,0%**	$62,5\%$	$71,2\%^{***}$	$87,1\%$	$93,4\%$	**32,3%**	
Del. 89+ days	**60,9%****	**91,6%***	$59,5\%^{***}$	**63,9%*****	**98%**	$34,5\%^{***}$	
Central	**63,2%****	$92,8\%^{*}$	**57,2%**	$56,6\%$	**98,4%**	**31,6%**	
North	$55,4\%$	**92,9%**	$55,4\%$	$64,3\%$	$96,4\%$	$21,4\%$	
South	$58,5\%$	**90,7%**	$61,2\%^{***}$	$66,3\%^{***}$	$97,8\%$	$35,9\%^{***}$	

BML forecast performance

Departing State	Binomial hit-rate (transition to): Current	Del. 30-89	Del. 89+	Default	Prepaid	Multinomial hit-rate	
Current	**72,6%*****	$97,5\%^{***}$	**97,7%*****	$99,4\%^{***}$	**77,0%*****	**72,1%*****	
Central	**71,6%*****	**98,1%*****	**97,9%*****	**99,4%****	$75,4\%^{***}$	**71,2%*****	g = 10
North	**77,2%*****	$97,2\%$	$97,4\%^{*}$	**99,5%*****	**82,0%*****	**76,4%*****	g = 100
South	**72,8%*****	$97,2\%^{***}$	$97,6\%^{***}$	$99,4\%^{***}$	**77,6%*****	**72,2%*****	g = 10
Del. 30-89 days	$49,0\%$	$65,0\%$	**68,5%*****	**88,0%**	**94,8%**	**34,3%***	
Central	**56,7%****	**69,0%**	$63,2\%$	**89,5%**	**96,5%**	**39,8%****	g = 1
North	$39,0\%$	$61,0\%$	**65,9%**	$90,2\%$	$97,6\%$	$34,1\%$	g = 2
South	$46,6\%$	**63,6%**	$71,2\%^{***}$	$87,1\%$	**93,7%**	$31,8\%$	g = 1
Del. 89+ days	$59,6\%$	$91,5\%^{*}$	**59,9%*****	$63,4\%^{***}$	$97,9\%$	**35,6%*****	
Central	$59,2\%^{***}$	**95,4%*****	$56,6\%$	$56,6\%$	$97,7\%^{**}$	$29,6\%$	g = 2
North	**66,1%*****	$85,7\%^{***}$	**58,9%**	**66,1%**	$96,4\%$	**39,3%**	g = 2
South	**59,1%*****	$90,1\%^{***}$	**61,5%**	**66,5%**	**98,1%*****	**38,1%*****	g = 0,1

(***): $\leq$ 1%, (**): $\leq$ 5%, (*): $\leq$ 10% - indicate significance levels over benchmark hit-rate. Values in **bold** indicate the best result between the ML model (top half) and the BML: geographically segmented model (bottom half).

For the mortgage DEL. 30-89 DAYS it is found that the addition of modeling the geographic segmentation allows for forecast accuracy to improve to a significance level of 10% over the benchmark. This result is mainly driven by the improvement of the the CENTRAL segment, which improved greatly with respect to the accuracy of the ML model. Also the forecasts for the mortgages departing the last state DEL. 89+ DAYS have improved slightly over those realized by the ML model. Although here it seem to be largely the result of relatively bad forecasts for the NORTH segment in the ML model.

Overall it can be stated that the modeling of the geographic segmentation, augmented with prior information where needed, increases out-of-sample forecasts slightly. For the aggregated segments per departing state, increase of a percentage point can be seen for the non-performing states DEL. 30-89 DAYS and DEL. 89+ DAYS. Whereas looking at individual segments the forecast accuracy can increase with multiple per-

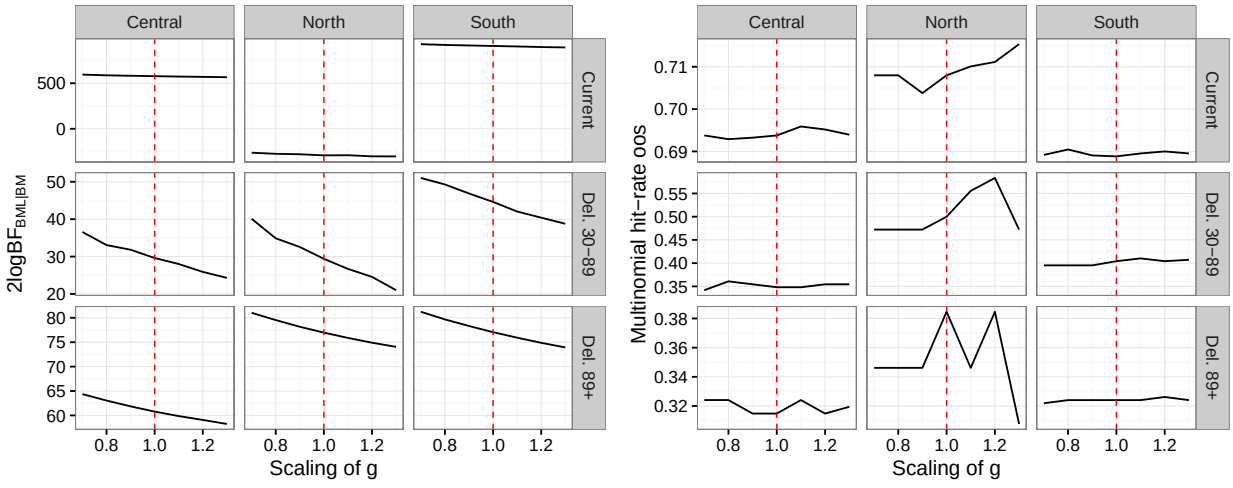
centage points. The largest gains in forecast accuracy are found to occur in the non-performing states where there are relatively fewer observations per segment. This is in-line with what was found in-sample where these non-performing states showed evidence for the segmented structure. For the departing state CURRENT we see slight improvements to forecasting accuracy. In-sample we did not find any evidence of an increase of fit for this departing state, but the increase in hit rate is also marginally better for the BML with geographic segmentation than for the ML model without segmentation.

With respect to the forecast results for vintage and K-means segmentation, the Tables 17 and 18 in Appendix A.1 show the results. As these results are in-line with what was already found, a deep exploration is left to reader, here I will state the main observations. Concentrating on multinomial hit-rate, in general for these two segmentation techniques we find what was also found with the geographic segmentation and in-sample results. Forecasts for CURRENT states tend to be marginally better, if at all, than the ML model. For vintage we see, just as in-sample, a slight improvement in the non-performing departing states DELINQUENCY 30-89 DAYS and DELINQUENCY 89+ DAYS with respect to forecast accuracy. Where with the K-means segmentation scheme we also see an improvement in these two non-performing departing states.

6.4 Sensitivity to prior specification

Due to the specification of the Bayesian multinomial logit model with a prior distribution, the question can arrise how sensitive the result are to the tuning parameter g. To analyze the effect of the choice of the hyper-parameter g on the in-sample Bayes factors and the out-of-sample multinomial hit-rate, the BML model was rerun several times for with scaled values of g. A range of plus and minus 30% was used to investigate the behavior of the BML model under slightly different circumstances. Figure 6.2 below summarizes the results of the sensitivity analysis.

Figure 6.2: Multinomial hit-rate and Bayes factors sensitivities



Log Bayes Factors (in-sample) and Multinomial hit-rate (out-of-sample) comparisons for scaling of hyper-parameter g around the respective chosen value.

When looking at the log Bayes factors it can be seen that the in-sample fit is sensitive to the specification of g. All Bayes factors, with the exception of those for the segments departing CURRENT exhibit a declining behavior. Altough it should be noted that the conclusions taken from the results discussed in Section 6.2 remain valid, as the Bayes factors do not chance sign from that which was reported previously.

Looking at the out-of-sample forecasting performance under various specifications of g, the right side of Figure 6.2, it can be seen that the geographical NORTH segments are more sensitive to changes in the prior than other segments. The NORTH segment is also the segment with the fewest observations (see Table 13). This can affect the forecasting accuracy in two ways. Firstly, the value of g is based on the in-sample forecast accuracy on a with held sub-sample. A smaller group to base the in-sample accuracy on results in larger parameter uncertainty when determining the optimal g. Secondly, fewer observations on which to calculate the hit-rate result in a larger variability when the state of a mortgage goes from being correctly forecast to incorrect, or vice-versa. It can be seen that for the segments CENTRAL and SOUTH the conclusions hold even under slight different specifications of g. For the NORTH segment we see that especially for the segments departing DEL. 30-89 DAYS and DEL. 89+ DAYS the results can vary by multiple percentage points. Do note however that at the specification of g used in the models neither of these two non-performing states showed a significant improvement over the benchmark.

7 Conclusion and discussion

In this research I aimed to better model the performance of a pool of mortgages. To this end I implemented the data-set provided by Freddie Mac covering 30-year fixed rate prime mortgages originated since 1999 and tracked monthly. I then followed broadly the work of Smith, Sanchez, and Lawrence (1996) and implemented a Markovian structure describing the evolution of mortgage performance between the states CURRENT, DEL. 30-89 DAYS, DEL. 89+ DAYS, DEFAULT and PREPAID in Section 4.1. To estimate the transition probabilities between states on a yearly basis, I implement a multinomial logit model, discussed in Section 4.2.

I expand on this by proposing a segmentation of the data to account for intuitive heterogeneity in the mortgages. I consider a geographic segmentation mirroring the known and documented cultural differences in the California and a segmentation by vintage of the mortgages. Additionally, I use a K-means algorithm to retrieve a data-driven segmentation of the mortgage pool. Although the addition of these segmentation schemes add granularity to the model, it also greatly increases the number of parameters that must be estimated with the same set of observations. This introduces parameter uncertainty and could negatively effect forecast accuracy. To tackle this I specify a Bayesian multinomial model in Section 4.5, where a prior is imposed on the parameters estimated in each segment based on the parameter estimates of the aggregated mortgage data per departing state. To be able to estimate the parameters I sample from the posterior distribution via the use of an adaptive Metropolis-Hastings algorithm discussed in Section 4.6. To achieve efficient sampling I propose some slight adjustments to the algorithm of Atchadé & Rosenthal (2005) and implement a multiplicative scaling procedure in the burn-in phase.

To analyze the multinomial forecasts I propose a multinomial decision rule in Section 5 which allows me to map the probability vector estimated for each mortgage to a unique state transition, whilst maintaining the implied properties on a mortgage pool level.

In analyzing the in-sample results I find that the more complex Bayesian multinomial model only has positive and very strong evidence for the departing states of DEL. 30-89 DAYS and DEL. 89+ DAYS respectively when applying a geographic or vintage segmentation. Whereas in the out-of-sample forecasts I find that accuracy is improved slightly for all departing states, but mainly for the non-performing departing states DEL. 30-89 DAYS and DEL. 89+ DAYS.

The modeling of the geographic segmentation shows evidence of a better fit than the more parsimonious counterpart in the cases of the non-performing departing states. The transitions departing from these states

were also found to be difficult to estimate with the standard multinomial logit model. That the expansion of the model with geographical segmentation yielded positive results can indicate that the way non-performing loans are handled can differ by location. Intuitively this makes sense as the manner in which a mortgage servicer handles non-performing loans can be expected to be under more scrutiny from regulators than performing loans. This is also emphasized by the lack of evidence that the performance of performing mortgages can be better estimated when accounting for geographical location, as healthy mortgages seem to evolve homogeneously in this regard.

A similar effect is found when segmenting by vintage of the mortgages. Where the age of a loan can indicate a different relation to the explanatory variables. One thing of note in this is that the age of the loan also says a lot about if the loan has a high loan-to-value ratio due to the fact that housing prices have declined since the inception of the mortgage. Additionally, after the crisis underwriting practices have become stricter and it can be expected that mortgages are granted to those better able to withstand economic and financial set-backs. As such the manner in which post-crisis mortgages react to falling in arrears is different that those pre-crisis.

For the data-driven K-means segmentation I find weaker results. This segmentation scheme does not appear to be able to capture the heterogeneity in the mortgage pool. As such its in-sample fit shows no evidence of improvement over a parsimonious non-segmented multinomial logit model. Initially this result is unexpected, as the implied segmentation of the mortgages made for an intuitive classification of risk and maturity of the loans. When considering the situation a bit more, there is no direct reason why loans with, for example, lower *credit score* or *unemp. rates* should react entirely differently to the explanatory variables. Especially as the factors used to create the K-mean segments already enter the model as explanatory variables.

Indeed it can be said that the addition of segmentation to the data adds significant explanatory value, but only in those cases where the choice of segmentation is not based on already considered (and accounted for) characteristics of the mortgages. In general, the application of segmentation and the implementation of a Bayesian multinomial logit was seen to provide some improvement in forecast accuracy. The main obstacle for this approach is to recognize when a sub-set of a considered set of mortgage requires separate modeling. By this I mean that the extra parametrization of the model will only pay-off if the segments are sufficiently different from the larger group.

As I mentioned in Section 3, exactly which factors are driving the development of mortgage performance will differ per country but also on a local regional level. The manner in which to estimate the effects of these drivers on mortgage performance however, as discussed in this paper, can be applied more generally over differing mortgage markets.

With respect to increasing the accuracy of mortgage performance I feel that the approach investigated in my research can serve as a basis for a more comprehensive performance model. Especially defining better, but not necessarily more, covariates that capture the dynamics of mortgage performance can yield increased accuracy. If indications of heterogeneity in a considered mortgage pool are found or motivated, and can be properly formulated, the segmentation technique in this paper can then be applied. A large, and possible understated result, of this thesis is the relatively good performance of the basic multinomial logit model with the identified explanatory variables. The explanatory variables considered and applied definitely added forecasting accuracy. As such I feel that one of the most important aspects of modeling mortgage performance is the identification and formulation of the proper drivers behind mortgage behaviour.

References

Albert, J. H., Chib, S., (1993), *Bayesian Analysis of Binary and Polychotomous Response Data* Journal of American Statistical Association, Vol. 88, No. 422 (June 1993), pages 669-679

Albert, I., Sophie, D., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., Rousseau, J., (2012), *Combining Expert Opinions in Prior Elicitation*, Bayesian Analysis, Volume 7, Num. 33, pages 503-532

Atchadé, Y. F., Rosenthal, J. S., (2005), *On adaptive Markov chain Monte Carlo algorithms*, Bernoulli, Volume 11, Num. 5, pages 815-828

Bayes, T., (1763), *An Essay towards solving a Problem in the Doctrine of Chances*, Philosophical Transactions of the Royal Society of London, Num. 53, pages 370-418

Campbell, S., Canlin, L., Im, J., (2014), *Measuring Agency MBS Market Liquidity with Transaction Data*, FEDS Notes [accessed July 13th 2015], `http://www.federalreserve.gov/econresdata/notes/feds-notes/2014/measuring-agency-mbs-market-liquidity-with-transaction-data-20140131.html`

Cox, J. C., Ingersoll, J. E., Ross, S. A., (1985), *A theory of the term structure of interest rates*, Econometrica, Vol. 53, No. 2, pages 394-402

Downing, C., Stanton, R., Wallace, N., (2005), *An Empirical Test of a Two-Factor Mortgage Valuation Model: How Much Do House Prices Matter?* Real Estate Economics, Vol. 33, No. 4 (2005), pages 681-710

Dunn, K., McConnell, J., (1981), *Valuation of GNMA mortgage-backed securities*, The Journal of Finance, Vol. 36 (1981), No. 3, pages 599-616

Geweke, J., (1992), *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Bayesian Statistics 4, Oxford: Oxford University Press, pages 169-193

Gilks, W. R., Wild, P., (1992), *Adaptive Rejection Sampling or Gibbs Sampling* Journal of the Royal Statistical Society, series C (Applied Statistics), Vol. 41, No. 2 (1992), pages 337-348

Goodarzi, A., Kohavi, R., Harmon, R., Senkut, A., (1998), *Loan prepayment modeling* KDD Workshop on Data Mining in Finance, edite by T. H. Hann and G. Nakhaezadeh 1998

Gneiting, T., Raftery, A. E., (2007), *Strictly proper scoring rules, prediction, and estiamtion* Journal of the American Statistical Association, 102:477, pages 359-378

Grimshaw, S. D., Alexander, W. P., (2001), *Markov Chain Models for Delinquency: Transition Matrix Estimation and Forecasting* Applied Stochastic Models in Business and Industry, Vol. 27, Num. 3, pages 267-279

Haario, H., Saksman, E., Tamminen, J., (2001), *An adaptive Metropolis algorithm* Bernoulli, Vol. 7, No. 2, 223-242

Hartigan, J. A., (1975), *Clustering Algorithms*, New York: Wiley

Hartigan, J. A., Wong, M. A., (1979), *A K-Means Clustering Algorithm*, Applied Statistics, Vol. 28, pages 100-108

Hastings, W. K., (1970), *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika, No. 57, pages 97-109

Hayres, L. S., Young, R., (2004), *Guide to Mortgage-Backed Securities*, Citigroup Fixed Income Research, November 3, 2004

Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K., (2005), *Econometric Methods with Applications in Business and Economics*, Oxford: Oxford University Press, ISBN 978-0-19-926801-6

Kang, P., Zenios, S. A., (1992), *Complete prepayment models for mortgage-backed securities*, Management Science, Vol. 38, No. 11, pages 1665-1685

Kass, R. E., Raftery, A. E., (1995), *Bayes Factors*, Journal of the American Statistical Association, Vol. 90, No. 430, pages 773-795

Kau, J. B., Keenen, D. C., (1995), *An Overview of the Option-Theoretic Pricing of Mortgages*, Journal of Housing Research, Vol. 6, No. 2, pages 217-244

Levin, A., Davidson, A., (2005), *Prepayment Risk- and Option-Adjusted Valuation of MBS*, The Journal of Portfolio Management, Volume 31, Num. 4, pages 73-85

Liang, T., Lin, J., (2014), *A two-stage segment and prediction model for mortgage prepayment prediction and management*, International Journal of Forecasting, 30(2014), pages 328-343

Milligan, G. W., Cooper, M. C., (1985), *An examination of procedures for determining the number of clusters in a data set*, Psychometrika, Volume 50, pages 159-179

Mian, A., Sufi, A., (2009), *The consequences of mortgage credit expansion: evidence from the U.S. mortgage default crisis*, The Quarterly Journal of Economics, Nov. (2009), pages 1449-1496

O'Brien, S. M., Dunson, D.B., (2004), *Bayesian Multivariate Logistic Regression*, Biometrics, Volume 60, Issue 3, pages 739-746

MacQueen, J. B., (1967), *Some methods for classification and analysis of multivariate observation*, Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability, pages 281-297

Pesaran, M. H., and Timmermann, A., (1992), *A simple parametric Test of Predictive Performance*, Journal of Business and Economic Statistics, Vol. 10(4), pp. 561-565

Polson, N. G., Scott, J. G., Windle, J., (2013), *Bayesian Inference for Logistic Models Using Pólya-Gamma latent Variables* Journal of American Statistical Association, Vol. 108, No. 504 (2013), pages 1339-1349

Popova, I., Popova, E., George, E. I., (2008), *Bayesian Forecasting of Prepayment Rates for Individual Pools of Mortgages* Bayesian Analysis, Vol 3, No. 2 (2008), pages 393-426

Schwarz, E. S., Torous, W. N., (1989), *Prepayment and the Valuation of Mortgage-Backed Securities*, The Journal of Finance, Volume 44, Num. 2, pages 375-392

Smith, L. D., Sanchez, S. M., Lawrence, E. C., (1996), *A Comprehensive Model for Managing Credit Risk on Home Mortgage Portfolios*, Decision Sciences, Volume 27, Num. 2, pages 292-317

Stanton, R., (1995), *Rational Prepayment and the Valuation of Mortgage-Backed Securities*, The Review of Financial Studies, Volume 8, Num. 3, pages 677-708

Sung, M., Soyer, R., Nhan, N., (2007), *Bayesian Analysis of Non-homogeneous Markov Chains: Application to mental health data* Statistics in Medicine, July 10, 26(15), pages 3000-17

Tibshirani, T., Walther, G., Hastie, T., (2001), *Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society B, Volume 63, part 2, pages 411-423

Zipkin, p., (1993), *Mortgages and Markov Chains: A Simplified Evaulation Model* Management Science, Vol. 39, No. 6 (Jun. 1993), pages 683-691

Maximum Loan Limits for 2010-Originated Mortgages, (2010), `http://www.fhfa.gov/DataTools/Downloads/pages/conforming-loan-limits.aspx`

A Appendix

A.1 Vintage & K-means forecasting accuracy

This appendix shows the results gotten when segmenting by vintage year and K-means algorithm for the out-of-sample forecast accuracy.

Table 17: BML out-of-sample accuracy, segmented by Vintage

ML forecast performance

Departing State	Binomial hit-rate (transition to): Current	Del. 30-89	Del. 89+	Default	Prepaid	Multinomial hit-rate
Current	*****71, 1%**	*****97, 4%**	*****97, 4%**	**99, 3%	*****75, 7%**	*****70, 4%**
Crisis	*****67, 7%**	**93, 9%**	**92, 4%**	**97, 7%**	***81, 1%	*****66, 2%**
Early	***71, 8%	*97, 5%	***98, 6%	99, 5%	***75, 2%	***71, 1%
Post-crisis	*****72, 4%**	***98, 6%	*****98, 9%**	***99, 6%	*****74, 4%**	*****71, 8%**
Pre-crisis	***69, 1%	**95, 8%**	95, 5%	99, 2%	***76, 7%	***68%
Del. 30-89 days	**51%**	67, 1%	**65, 2%**	87%	**94, 5%**	32, 2%
Crisis	37, 1%	68%	**64, 9%**	86, 6%	96, 9%	27, 8%
Early	61, 5%	**69, 2%**	**61, 5%**	92, 3%	69, 2%	30, 8%
Post-crisis	****57%**	69, 1%	67, 8%	**84, 6%**	**91, 9%**	33, 6%
Pre-crisis	50, 6%	**66, 4%**	63, 2%	88, 1%	95, 3%	31, 8%
Del. 89+ days	****60, 9%**	*91, 6%	***59, 5%	*****63, 9%**	**98%**	***34, 5%
Crisis	**58, 1%**	***92, 5%**	****60, 6%**	63, 1%	98, 9%	31, 2%
Early	55%	**90%**	50%	65%	**80%**	***45%**
Post-crisis	**61, 1%**	90, 7%	57, 5%	**62, 8%**	98, 2%	*34, 1%
Pre-crisis	**60, 3%**	91, 5%	****59, 7%**	****64, 5%**	97, 4%	***35, 1%

BML forecast performance

Departing State	Binomial hit-rate (transition to): Current	Del. 30-89	Del. 89+	Default	Prepaid	Multinomial hit-rate	
Current	70, 9%	97, 3%	97, 3%	**99, 4%**	75, 6%	***69, 9%	
Crisis	**66%	92, 1%	91%	97, 4%	*****82, 3%**	*63, 7%	g=100
Early	*****74, 3%**	****97, 9%**	*****99, 1%**	99, 5%	*****77, 5%**	*****74, 2%**	g=10
Post-crisis	***71, 9%	*****98, 7%**	***98, 6%	***99, 6%	***74, 1%	***71, 3%	g=5
Pre-crisis	*****69, 6%**	95, 7%	**95, 9%**	**99, 3%**	*****76, 8%**	*****68, 3%**	g=10
Del. 30-89 days	50, 3%	**67, 2%**	64, 5%	**87, 3%**	94, 1%	**32, 4%**	
Crisis	**41, 2%**	68%	61, 9%	**89, 7%**	****99%**	27, 8%	g=10
Early	61, 5%	61, 5%	38, 5%	92, 3%	**84, 6%**	30, 8%	g=1
Post-crisis	54, 4%	**71, 8%**	67, 8%	82, 6%	89, 3%	**36, 2%**	g=1
Pre-crisis	50, 6%	65, 1%	**64, 8%**	**88, 7%**	95, 3%	**32, 1%**	g=0,01
Del. 89+ days	59, 2%	**91, 8%**	**59, 6%**	62, 6%	97, 7%	*****35, 5%**	
Crisis	**56, 3%	***91, 8%	58, 4%	63, 1%	***98, 9%	**33, 7%**	g=10
Early	***65%**	*85%	**60%**	**70%**	75%	30%	g=1
Post-crisis	***60, 6%	*****92%**	**61, 9%**	59, 3%	*****98, 7%**	****38, 9%**	g=1
Pre-crisis	***60, 1%	*****91, 9%**	59%	63, 6%	**97, 4%	35, 1%	g=1

(***): $\leq$ 1%, (**): $\leq$ 5%, (*): $\leq$ 10% - significance levels. Values in **bold** indicate the best result between the ML model (top half) and the BML: vintage segmented model (bottom half).

Table 18: BML out-of-sample accuracy, K-means segmentation

ML forecast performance

Departing State	Binomial hit-rate (transition to): Current	Del. 30-89	Del. 89+	Default	Prepaid	Multinomial hit-rate	
Current	*****71, 1%**	*****97, 4%**	***97, 4%	****99, 3%**	*****75, 7%**	*****70, 4%**	
Normal	*****67, 5%**	**96, 2%**	**95%**	**98, 8%**	***75, 3%	*****66, 2%**	
Low Risk	*****72, 2%**	*****98, 8%**	***99, 5%	***99, 7%	*****73, 4%**	*****71, 7%**	
Mature	***66, 8%	***97, 8%	***98, 4%	***99, 8%	***69, 8%	***66, 4%	
High Risk	***75, 8%	94, 9%	94%	**98, 4%**	***86, 4%	***74, 6%	
Del. 30-89 days	51%	**67, 1%**	**65, 2%**	87%	94, 5%	32, 2%	
Normal	43, 9%	**71, 1%**	66, 3%	**84%**	95, 2%	31%	
Low Risk	54, 1%	**75, 4%**	****73, 8%**	90, 2%	91, 8%	**42, 6%	
Mature	**54, 8%**	58, 9%	**66, 9%**	***95, 2%	87, 1%	**30, 6%**	
High Risk	**48, 3%**	**68, 3%**	62%	82, 9%	*97, 6%	27, 8%	
Del. 89+ days	****60, 9%**	*91, 6%	***59, 5%	*****63, 9%**	98%	***34, 5%	
Normal	*60, 9%	91, 3%	*****60%**	***64, 9%	**98, 4%**	***36%	
Low Risk	***71%**	90, 3%	61, 3%	61, 3%	100%	38, 7%	
Mature	49, 5%	**87, 2%**	**59, 6%**	***74, 3%	90, 8%	****37, 6%**	
High Risk	**59, 6%**	****92, 9%**	**58, 6%	55, 6%	*99, 5%	30, 2%	

BML forecast performance

Departing State	Binomial hit-rate (transition to): Current	Del. 30-89	Del. 89+	Default	Prepaid	Multinomial hit-rate	
Current	70, 5%	97, 2%	97, 4%	99, 2%	75, 3%	***69, 8%	
Normal	***66, 7%	96%	93, 6%	98, 7%	*****75, 7%**	***65, 2%	g=5
Low Risk	***71, 1%	***98, 3%	*****99, 8%**	***99, 7%	***72, 7%	***70, 8%	g=100
Mature	*****67, 2%**	*****98%**	*****98, 6%**	***99, 8%	*****70, 1%**	*****67%**	g=100
High Risk	*****76, 5%**	**95%**	**94, 5%**	97, 8%	**86, 7%**	*****75, 2%**	g=10
Del. 30-89 days	**51, 1%**	66, 4%	64, 8%	**87, 5%**	**95, 3%**	**33, 1%**	
Normal	**49, 2%**	69%	66, 3%	83, 4%	**95, 7%**	**34, 2%**	g=5
Low Risk	****62, 3%**	67, 2%	*72, 1%	**91, 8%**	**93, 4%**	****44, 3%**	g=5
Mature	54%	**59, 7%**	62, 1%	*****98, 4%**	**89, 5%**	29%	g=5
High Risk	47, 8%	67, 8%	**62, 9%**	**83, 4%**	*****99%**	**31, 2%**	g=1
Del. 89+ days	59%	**91, 7%**	**59, 9%**	62, 5%	98%	*****35, 5%**	
Normal	*****62, 4%**	*****92, 9%**	59, 3%	64, 9%	***98, 2%	*****38, 7%**	g=5
Low Risk	61, 3%	***90, 3%	**67, 7%**	**64, 5%**	100%	**41, 9%**	g=1
Mature	49, 5%	***85, 3%	58, 7%	**75, 2%**	90, 8%	34, 9%	g=2
High Risk	***57, 6%	***92, 1%	**60, 2%**	**56, 1%**	***99, 5%	**31, 5%**	g=1

(***): $\leq$ 1%, (**): $\leq$ 5%, (*): $\leq$ 10% - indicate significance levels over benchmark hit-rate. Values in **bold** indicate the best result between the ML model (top half) and the BML: K-means segmented model (bottom half).

A.2 Zip-codes & Counties

Table 19: Distribution of 3-digit zip codes over segments

3-digit code	Major counties	Segmentation
90000	Los Angeles County	South
90200	Los Angeles County	South
90300	Los Angeles County	South
90400	Los Angeles County	South
90500	Los Angeles County	South
90600	Los Angeles County	South
90700	Los Angeles County	South
90800	Los Angeles County	South
91000	Los Angeles County	South
91100	Los Angeles County	South
91200	Los Angeles County	South
91300	Los Angeles County	South
91400	Los Angeles County	South
91500	Los Angeles County	South
91600	Los Angeles County	South
91700	Los Angeles County	South
91800	Los Angeles County	South
91900	San Diego County	South
92000	San Diego County	South
92100	San Diego County	South
92200	64% Riverside, 27% Imperial County, 9% San Bernadino	South
92300	San Bernadino County	South
92400	San Bernadino County	South
92500	Riverside County	South
92600	Orange County	South
92700	Orange County	South
92800	80% Orange County, 20% Riverside County	South
93000	Venture County	South
93100	Santa Barbara	South
93200	54% Tulare, 20% Kings, 26% Kern County	South
93300	Kern County	South
93400	55% San Luis Obispo, 45% Santa Barbera	South
93500	25% Kern, 75% Los Angeles County	South
93600	66% Fresno, 33% Maderna	Central
93700	Fresno County	Central
93800	Fresno County	Central
93900	Monterey County	Central
94000	70% San Mateo, 30% Santa Clara County	Central
94100	San Fransisco County	Central
94300	Santa Clara County	Central
94400	San Mateo County	Central
94500	50% Alemeda, 50% Contra Costa County	Central
94600	Alemeda County	Central
94700	Alemeda County	Central
94800	Contra Costa	Central
94900	67% Marin, 33% Sonoma County	Central
95000	67% Santa Clara, 33% Santa Cruz County	Central
95100	Santa Clara County	Central
95200	San Joaquin County	Central
95300	20% Merced, 25% San Joanquin, 55% Stanislaus County	Central
95400	75% Sonoma, 10% Lake, 15% Mendocino County	North
95500	85% Humboldt, 15% Del Norte County	North
95600	55% Sacremento, 25% Placer, 20% Yolo County	Central
95700	45% Placer, 35% Sacramento, 20% Eldorado County	Central
95800	Sacramento County	Central
95900	45% Butte, 20% Stutter, 20% Nevada, 15% Yuba County	North
96000	60% Shasta, 15% Siskiyou, 25% Tehema County	North
96100	30% El Dorado, 30% Lassen, 20% Nevada, 20% Placer County	North

A.3 Federal Home Loan Mortgage Corporation data-set

Table 20: Freddie Mac single family 30-year fixed rate data-set (1)

Freddie-Mac field name and description	Field values	Explanation	Monthly/ Origination
LOAN SEQUENCE NUMBER Unique sequence number per loan.	F1YYQn#		M/O
MONTHLY REPORTING PERIOD Reporting month.	YYYYMM		M
CURRENT ACTUAL UPB Unpaid balance: interest bearing + non-interest bearing.	#	In dollars nominal	M
CURRENT DELINQUENCY STATUS A value corresponding to the number of months a borrower is delinquent.	#	in months	M
LOAN AGE Number of full months since note origination month.	#	in months	M
REMAINING MONTHS TO LEGAL MATURITY The remaining months to the mortgage maturity date.	#	in months	M
ZERO BALANCE CODE A code indicating the reason the loan's balance was reduced to zero.	1 3 6 9	Prepaid or matured Foreclosure Repurchase prior to prop. Disp. REO disposition	M
CURRENT INTEREST RATE Current interest rate on the mortgage note, taking into account any modifications.	#	Percentage	M
CREDIT SCORE Third party credit score of borrowers creditworthiness.	#		O
FIRST TIME HOMEBUYER FLAG Indicates whether the borrower, or one of the borrowers, is a first time home buyer.	Y N space(1)	Yes No Unknown	O
METROPOLITAN STATISTICAL AREA (MSA) 2010 Metropolitan Statistical Area.	#		O
MORTGAGE INSURANCE PERCENTAGE Percentage of loss coverage on the loan, at the time of FM's purchase of the mortgage loan that a mortgage insurer is providing to cover losses.	#	Percentage	O
NUMBER OF UNTIS Denotes whether the mortgage is a 1,2,3 or 4 unit property.	1 2 3 4 space(1)	one unit two unit three unit four unit Unknown	O
OCCUPANCY STATUS Denotes whether the mortgage type is owner occupied, second home, or investment property.	O I S NA	Owner Occupied Investment Property Second Home Unknown	O

Table 21: Freddie Mac single family 30-year fixed rate data-set (2)

Freddie Mac field name and description	**Field values**	**Explanation**	**Monthly/ Origination**
ORIGINAL COMBINED LOAN-TO-VALUE (CLTV) Sum of all mortgage loan amounts divided by property value.	0%-200% space(3)	Percentage Unknown	O
ORIGINAL DEBT-TO-INCOME (DTI) RATIO Sum of borrowers monthly debt payments to income.	0% - 65% space(3) Null	DTI DTI >65% Unknown	O
ORIGINAL UPB The unpaid balance on the mortgage note date.	#	In dollars nominal	O
ORIGINAL LOAN-TO-VALUE (LTV) Original mortgage loan amount divided by (the lesser of) the purchase price and the property's appraised value.	6% - 105% space(3)	Unknown, <6% or >105%	O
CHANNEL Indicates the channel by which the loan was originated.	R B C T	Retail Broker Correspondent TPO/ Not Specified	O
PREPAYMENT PENALTY MORTGAGE (PPM) FLAG Flag indicating if the mortgage borrower is subject (or has been) to a penalty with respect to pre-payment of principal.	Y N NA	Prepayment Penalty not Prepayment penalty unknown	O
PROPERTY TYPE Denotes the type of property secured by the mortgage.	CO CP LH MH PU SF space(2)	Condo Co-Op Leasehold Manufactured Housing PUD 1-4 Free Simple Unknown	O
POSTAL CODE The postal code for the location of the mortgage property. Only the first 3 digits of the postal code are available.	###00		O
LOAN PURPOSE Indicates the purpose of the mortgage loan.	C N P	Cash-Out Refinance No Cash-out refinance Purchase	O
NUMBER OF BORROWERS The number of borrowers who are obligated to repay the mortgage note.	1 2 space(2)	1 borrower more than 1 borrower Unknown	O
SELLER NAME Entity acting as seller of the mortgage to Freddie Mac at the time of acquisition.	name		O
SERVICER NAME Entity acting in its capacity as the servicer of mortgages to Freddie Mac as of the last period of activity of the mortgage in the data set.	name		O