Cash flow modelling for Residential Mortgage Backed Securities:

a survival analysis approach

Master thesis Applied Mathematics

Roxanne Busschers

September 2, 2011

UNIVERSITY OF TWENTE.


NIBC

This is a dissertation submitted for the Master Applied Mathematics
(Financial Engineering).
University of Twente, Enschede, The Netherlands.
Department of Electrical Engineering, Mathematics and Computer Science.

Master Thesis	
Title:	Cash flow modelling for Residential Mortgage Backed Securities: a survival analysis approach
Host organisation:	NIBC Bank N.V.
Research period:	February 2011 - July 2011
Author	
Name:	Roxanne Busschers
Student number:	s0128104
University:	University of Twente
Master degree program:	Applied Mathematics
Track:	Financial Engineering
Contact:	r.a.busschers@alumnus.utwente.nl
Supervisor Committee	
Supervisors University of Twente:	Prof. dr. A. Bagchi Dr. J. Krystul
Supervisor NIBC:	A.J. Broekhuizen

UNIVERSITY OF TWENTE.

Preface

This dissertation is part of my final project for the Master Applied Mathematics, specialisation Financial Engineering, at the University of Twente in Enschede. After my bachelor in Industrial Engineering and Management, I decided that this Master was going to be my next challenge and I have derived pleasure from every step of it. In February of this year I started working on my final project at NIBC Bank N.V. in The Hague. I worked there on a challenging assignment from practice, while at the same time experiencing the dynamics of the business world, which I have really enjoyed.

I want to thank Ton Broehuizen as my supervisor at NIBC for his ideas and guidance. Also, I like to thank my direct colleagues Dennis Hendriksen, Egbert Schimmel and Bálint Vágvölgyi for their ideas, comments, support and the pleasant working environment. My special thanks goes to Peter Kuijpers for all the time he took to discuss with me mortgage data and models. Finally, there are many more people at the bank who have helped me on several issues, in the completion of my project. Although, I cannot name them all here, I am very grateful to them.

From the university my project was supervised by Prof. Bagchi and dr. Krystul. Their guidance throughout the project helped me a lot and I like to thank them for that.

Roxanne Busschers.

Abstract

This thesis describes the research into modelling cash flows for Residential Mortgage Backed Securities (RMBS). RMBS notes are secured by proceeds, interest and principal payments, of the underlying mortgage pool. A transaction is divided into several classes of notes with different risk profiles, though they all reference to the same underlying assets.

The quality or creditworthiness of an RMBS transaction is assessed by credit rating agencies. During the credit crisis substantial losses were suffered on several RMBS notes, sometimes up to the most senior ones. In response, the rating agencies downgraded a lot of RMBS transactions, and more importantly the market questioned the ability of the rating agencies to assess the quality of structured credits. As a consequence pricing RMBS notes became very subjective. This forces investors to develop their own pricing models instead of relying on rating agencies. Finally, regulatory supervisors have reacted by requesting more transparency from issuers, resulting in the obligation for issuers to make available to investors detailed loan-level data on the underlying mortgage pool. The new regulations gave rise to research on how to purposefully apply loan-level data to consistently and arbitrage free value an RMBS note.

In this thesis we develop a model based on loan-level data to forecast the cash flows to the noteholders. This model has a stochastic part, the cash flows from the mortgage pool, and a deterministic part, the allocation of these cash flows to the noteholders established by the transaction structure. Besides interest payments, default and early repayment are determinants of

the size and timing of cash flows from the underlying mortgage pool. In this research, both the default and early repayment model are based on survival analysis, which allows for the estimation of month-to-month default and early repayment probabilities at a mortgage level. The Cox proportional hazards model adopted is able to incorporate both mortgage specific variables and time-varying covariates relating to the macro-economy. Since both default and early repayment can cause a mortgage to be terminated before maturity, these causes are termed 'competing risks'. In this paper we will extend the Cox model such that it explicitly accounts for the competing risk setting. We find that the probability of default for a mortgage is higher if:

- the ratio of loan to foreclosure value is higher;
- the borrower has a registered negative credit history;
- the ratio of main income to total income associated with the loan is higher;
- there is only one registered borrower;
- the income of the borrower is not disclosed to the lender, but to an intermediary.

For early repayment, we find that the probability of occurrence for a mortgage is higher if:

- the ratio of loan to foreclosure value is higher;
- the applicant is younger;
- the total income of the borrower(s) is lower;
- the 3-months Euribor is higher;
- it is an interest reset date;
- the refinancing incentive is higher.

We obtain a method to estimate the month-to-month default and early repayment probabilities for a specific mortgage with certain characteristics and age. Monte Carlo simulation is used to compute different realisations of default and early repayment for the underlying mortgage pool over the maturity of the RMBS. Finally, the deterministic structure of the notes allows us to derive the corresponding discounted cash flows to the noteholders and estimate a profit distribution for an RMBS note.
The research resulted in a tool for NIBC to assess the quality of a mortgage pool and employ this information to arbitrage free value a corresponding RMBS note.

Contents

List of Illustrations

Figures

Tables

Acronyms and abbreviations

Abbreviation	full name	description
ABS	Asset Backed Security	
BKR	Bureau Krediet Registratie	Adverse credit history is registered by BKR
bps	basispoints	equal to one-hundredth of a percentage point
CF	Cash flow	
CIF	Cumulative Incidence Function	probability of failing from a specific cause before time t
df	discount factor	
ER	Early Repayment	
FORD	First Optional Redemption Date	first date at which the issuer can redeem all notes of an RMBS
LGD	Loss Given Default	
NHG	Nederlandse Hypotheek Garantie	Dutch mortgage guarantee system
NIBC	Nederlandse Investerings Bank Capital	
NPV	Net Present Value	
PD	Probability of Default	
PDL	Principal Deficiency Ledger	
RMBS	Residential Mortgage Backed Security	
SPV	Special Purpose Vehicle	

Chapter 1

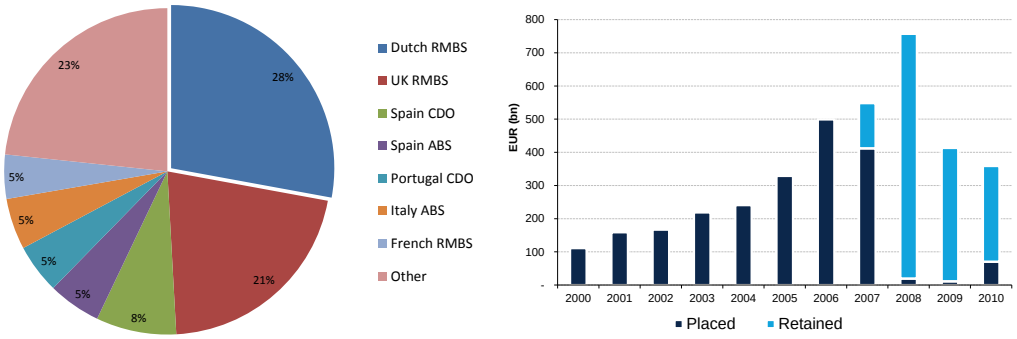
Introduction

In this first chapter we will give an introduction to the subject of Residential Mortgage Backed Securities and define the scope of the performed research, while at the same time motivating the reason of this research. The second section will describe the organization of this thesis.

1.1 Scope and motivation of the research

Residential Mortgage Backed Security (RMBS) notes are secured by proceeds, interest and principal payments, of the underlying mortgage pool. A transaction is divided into several classes of notes with different risk profiles, though they all reference to the same underlying assets. The different risk profiles are due to the transaction structure, which is generally quite complex but can, in short, be summarised as follows: income from interest or principal repayment is in general first distributed to the most senior ranking class. With losses, due to missed interest and principal payment, it works the other way around. These are first allocated to the junior class of the transaction. In other words, the more senior a class is, the less risk it bears of missing interest payments and losing part of the principal. Consequently more junior classes are offered a higher return to compensate for the higher risks investors bear.

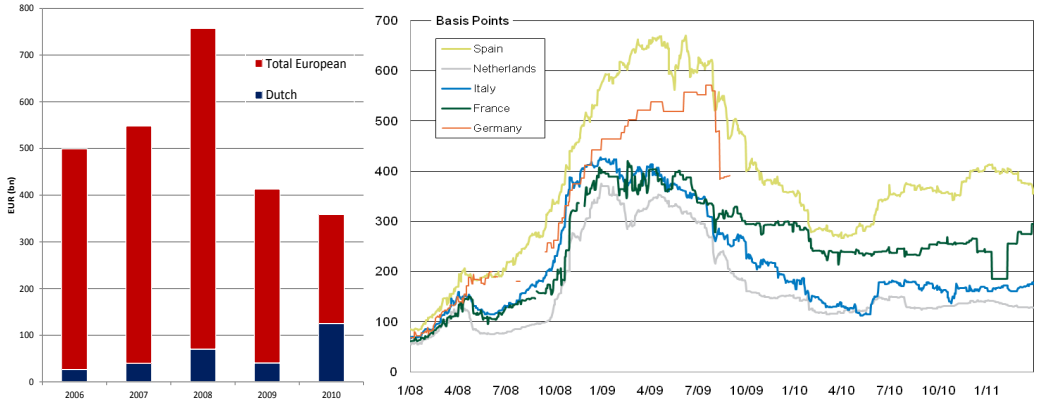
To gain an idea of the scope and importance of pricing adequately (Dutch) RMBS notes especially since the credit crisis, we will give a brief overview of the market for this financial product. In the first quartile of 2011 €31.9 billion of securitised Dutch RMBS transactions were issued, which amounts to almost 47% of total European RMBS issued and 28% of total European issued Asset Backed Securities (ABS). Figure 1.1a gives a graphical overview of European issuance of ABS in the first quartile of 2011. Figure 1.1b shows the absolute value of European supply of RMBS, publicly sold and retained by the issuer, in the years 2000 till 2010. From this figure we can clearly see the impact of the crisis in the years 2007 and later, when the market for all ABS collapsed.



(a) European ABS issuance first quartile 2011 (b) European supply of RMBS

Figure 1.1: Overview of European ABS market. Source: Association for Financial Markets in Europe (2011).

For the Dutch RMBS market there was € 289 billion outstanding collateral at the end of the first quartile of 2011, which amounts to 91% of the total ABS market in the Netherlands, indicating the significant size and relevance of RMBS transactions within the Dutch ABS market. It also accounts for 22.5% of total European outstanding collateral in RMBS transactions, which makes Dutch RMBS notes a significant contributor to the European market. Figure 1.2a shows the total size of the Dutch mortgage pools underlying the issuance over the last few years in absolute value and as a fraction of



(a) Dutch supply of RMBS (b) European spread on AAA-rated RMBS notes

Figure 1.2: Overview of Dutch RMBS market. Source: Association for Financial Markets in Europe (2011).

European issuance of RMBS notes. Finally, figure 1.2b displays how the spread in basispoints (equals one-hundreth of a percentage point) on AAA-rated RMBS notes have evolved over the last few years for a few European countries, including the Netherlands. The spread offered to investors is an indication of the risk the market anticipates for the financial product. Spread on Dutch RMBS have been relatively low, indicating that these products are still a relatively safe investment.

The quality or creditworthiness of an RMBS transaction is assessed by credit rating agencies (Moody's, Fitch and S&P). During the credit crisis substantial losses were suffered on several RMBS notes, sometimes up to the most senior ones. In response the rating agencies downgraded a lot of RMBS transactions, and more importantly the market questioned the ability of the rating agencies to assess the quality of structured credits. As a consequence pricing RMBS transactions became very subjective. This forced investors to develop their own pricing models instead of relying on rating agencies. Finally, regulatory supervisors have reacted in requesting more transparency from issuers. Therefore, issuers of new RMBS transactions are obliged to provide loan-level data in the near future. This implies that issuers of RMBS notes have to deliver to investors a large datafile containing a number of pre-specified mort-

gage characteristics of all securitisated residential mortgage loans, and also present investors with a frequent update of this file. Note that in this paper we will simply speak of a mortgage loan or a mortgage when referring to a residential mortgage loan.

The new regulations give rise to research on how to purposefully use this detailed loan-level data to be able to consistently and arbitrage free value an RMBS note. In this thesis we will develop a model based on loan-level data to forecast the cash flows to the noteholders. This model has a stochastic part, the cash flows originating from the mortgage pool, and a deterministic part, the allocation of these cash flows to the noteholders established by the transaction structure. The unknown cash flows from the mortgages cause the risk to the investor. The size and timing of these cash flows are unknown due to three main reasons:

- interest payments of an individual mortgage will change at an interest reset date. Since the underlying mortgage pool of an RMBS can consist of thousands of mortgages, it is impossible to know the resulting interest cash flows. This risk is often mitigated in the structure by an interest rate swap. The next chapter discusses this in more detail.

- a borrower could default on his mortgage, in that case it might happen that the proceeds of selling the house will not cover the entire outstanding loan. If the borrower cannot cover for the remaining amount, a loss might be incurred by the noteholders. In first instance these losses will be allocated to the most junior notes.

- a borrower could repay his mortgage before maturity, for example when he decides to move or refinance his mortgage elsewhere. The proceeds of this repayment are in general sequentially distributed to the most senior notes.

It is since long recognized that the probabilities of default and early repayment may vary over the duration of the loan. Therefore we need to develop a dynamic model which reflects the particular structure of the mortgage as well

as the economic changes that may occur during the outstanding period of the loan. To this end, both the default and early repayment model are based on survival analysis, which allows for the estimation of month-to-month default and early repayment probabilities at a mortgage level. The Cox proportional hazards model adopted is able to incorporate both mortgage specific variables and time-varying covariates relating to the macro-economy. Since both default and early repayment can cause a mortgage to be terminated before maturity, these causes are termed 'competing risks'. In this thesis we will extend the Cox model such that it explicitly accounts for the competing risk setting. Monte Carlo simulation is used to compute different realisations of default and early repayment for the underlying mortgage pool over the maturity of the RMBS. The deterministic structure of the notes allows us to derive the corresponding discounted cash flows to the noteholders and estimate a profit distribution for an RMBS note.

Rating agencies simply use a rating scale to express the risk in a bond from a loss perspective. Thus, a AAA rating estimates the risk of a loss (i.e. missed payment) as less then 0.01%. However, the value of a note also depends upon the interest rate used in discounting and can therefore change without a missed payment. The model we develop accounts for the uncertainty in size and timing of cash flows and therefore will give the complete distribution function of the value of a note. In this respect it distincts itself from the approach of the rating agencies as it can be used for valuation as well as risk management, indicating the uncertainty of the expected cash flows.

Although NIBC is an active originator in the Dutch market of RMBS issues, it is also an investor in RMBS notes. Besides investments in RMBS notes issued by other firms, including foreign banks, NIBC also holds a share of the RMBS notes it issued itself. Reasons for investing in RMBS notes issued in-house are, next to profitability, a regulatory obligation to retain part of the issued notes and the inability to sell (all) non-senior notes since the outburst of the credit crisis. For this thesis we take the perspective of NIBC as an investor in RMBS notes. While since the credit crisis, investors

only carefully invest in the most senior notes, an issuer will, for the reasons previously mentioned, also have riskier notes in its portfolio. Therefore, our interest will not merely be in the most senior notes but in all notes of an RMBS. Our focus will primarily be on notes issued by NIBC, although we assume that our model is also applicable to other Dutch RMBS notes.

1.2 Organization of the thesis

The organization of this thesis is as follows: in chapter two we will give an overview of Residential Mortgage Backed Securities and their characteristics. Chapter three describes the framework of the pricing tool for RMBS notes which we will develop in this project. It discusses in general the modelling of unknown cash flows and the simulation process combining the stochastic cash flows of the mortgage pool with the deterministic structure of an RMBS. Chapter four gives an overview of the existing literature on prepayment and default models as well as on how to model the incurred loss when a default occurs. Chapter five introduces survival analysis, which we will apply in this project to estimate the probability of default and early repayment of a mortgage. By applying a competing risk model we also explicitly account for the fact that a mortgage may be either terminated by default or by early repayment. In chapter six we discuss the mathematical details of our model, such as the formulas necessary to estimate the parameters of the model. Chapter seven then describes the characteristics of the data set we use to obtain the models. This chapter also discusses the model development steps. Chapter eight reports the results of the default and early repayment model for mortgages as well as the obtained results for a specific RMBS transaction. Finally, the last chapter draws conclusions and gives recommendations on further research on the model and improvement of the developed valuation tool.

Chapter 2

Overview of Residential Mortgage Backed Securities

Residential Mortgage Backed Securities (RMBS) are financial securities with mortgage loans as the underlying asset. Although there might be significant differences between RMBS transactions we will describe in this chapter the general characteristics.

2.1 Securitisation process

The process of creating a Residential Mortgage Backed Security is called securitisation. This process goes as follows: the originator (usually a bank or an insurance company) has a portfolio of residential mortgages, called the collateral pool, on its balance sheet and sells them to a so-called *Special Purpose Vehicle* (SPV). An SPV is a legally independent entity, which is most often created by the originator and has as a sole purpose the securitisation process. The arrangement has the effect of insulating investors from the credit risk of the originator. For mortgage originators, there are several reasons to issue mortgage backed securities, the most important are:

- transform relatively illiquid assets (mortgages) into liquid and tradable market instruments (notes);

- the originator may obtain funding at lower cost by securitisation than by borrowing directly in the capital markets;
- it allows the issuer to diversify his financing sources, by offering alternatives to more traditional forms of debt and equity financing;
- removing assets from the balance sheet, which can help to improve various financial ratios and reduce the exposure risk.

The SPV raises funds by issuing notes to investors structured as multiple classes, called *tranches*. These tranches have different seniority, ranging from most senior (typically rated AAA) to equity (typically unrated).
The fact that different tranches have different risk profiles, though they all reference to the same underlying assets, is based on the transaction structure. This enables investors to satisfy their individual appetites and needs. Assuming that the notes are sold at par (the face value) the equity tranches will, due to the higher risk, earn a higher return. This return will often consist of a floating part and a spread, for example 3-months Euribor + x basis points. Figure 2.1 depicts the general structure of a typical RMBS by a clarifying example.

2.2 Principal waterfall

The underlying mortgage pool generates interest and principal payments which are distributed via the interest and principal waterfall. The source for the principal waterfall consists besides principal repayments of foreclosure proceeds. Principal can be paid sequential or on a pro rata basis. If the principal is paid on a sequential basis, the senior notes are at the top of the waterfall and only after the senior notes are fully redeemed, principal payment is distributed to the mezzanine notes. For principal waterfalls in which principal is distributed on a pro rata basis, the transaction often incorporates triggers to protect senior notes. Such a transaction can be triggered, for example, by a high level of defaults, after which a switch is made from

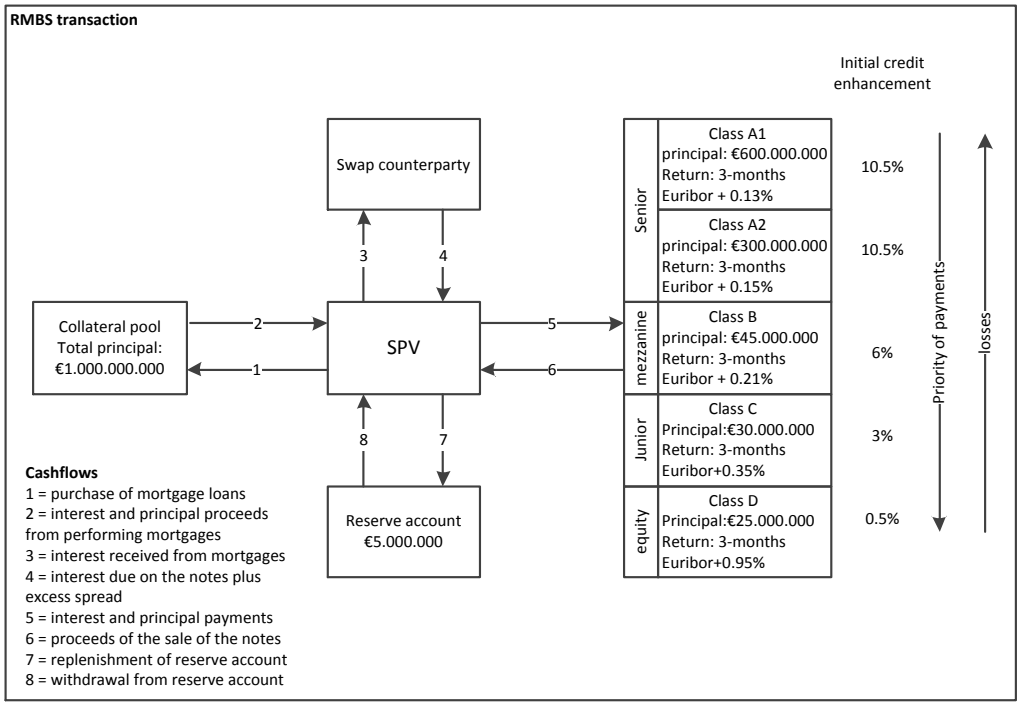


Figure 2.1: RMBS example

pro rata payments to sequential payments. However, there are many other structures of principal waterfalls possible.

If the portfolio of assets starts to experience default losses, these losses are first allocated to the equity tranche by reducing the outstanding amount of this tranche. This affects both the payment of principal as the payment of interest, since interest is paid over the remaining outstanding amount in the tranche.

2.3 Credit enhancement

Credit enhancement is the percentage loss that can be incurred on the mortgages before one Euro of loss is incurred on a particular note, see figure 2.1. There are several ways in which the structure can increase it. Credit enhancement techniques can be broadly divided into four categories, which we

will shortly discuss.

- Subordination is the first line of defence in an RMBS transaction. A tranche will only start to experience losses after the tranches subordinate to it are completely written off. For the most senior notes this implies all other notes, for mezzanine notes this implies the junior and equity tranche.

- A reserve account can be created to reimburse the SPV for losses up to the amount credited to the reserve account.

- Excess spread can be seen as the 'fat' in a structure. If the underlying mortgage pool yields on average a higher interest than the average interest on the notes (minus certain costs) some 'excess' stays in the SPV. When it is incorporated in the structure, it will absorb the first losses. Another use for excess spread can be to create and maintain the reserve account.

- Overcollateralisation ensures that the underlying collateral pool has a face value higher than the issued notes. Because the SPV owns more assets than it has debt with the noteholders, there is some extra certainty for these noteholders.

2.4 Interest swap and interest waterfall

The interest waterfall establishes the distribution of interest received from the mortgage pool. In most RMBS transactions the proceeds of an interest rate swap are used for interest payments to the noteholders. RMBS notes normally pay floating rates, whilst the mortgage collateral consist of mortgages with fixed and floating interest. To hedge the resulting interest rate risk an RMBS often incorporates an interest rate swap. This is another feature protecting investors from risks other than those arising from the mortgage pool. In general the SPV pays the swap counter party:

- Scheduled interest on the mortgages;
- plus prepayment penalties,

and the swap counter party pays the SPV:

- scheduled interest on the notes;
- plus excess spread if applicable.

While in the most simple case noteholders will always receive the interest payments, in more complex situations payment of interest is done on a sequential basis where the senior notes are at the top of the waterfall. The interest waterfall is subject to changes when it incorporates triggers that are activated. In these instances, the interest proceeds that would normally go to the mezzanine and equity tranches could be redirected to pay down the senior notes.

Finally, we mention so-called Principal Deficiency Ledgers (PDL's). When excess spread comes from the interest rate swap it is most often used through the PDL's to (partly) make up for incurred losses. There is a separate PDL for each tranche and it records any shortfall that would occur in repayment of the outstanding notes. Thus, when due to a loss the size of a tranche is reduced, the same amount is written to the corresponding PDL. The excess spread is than used in order of seniority to reduce the PDL and thereby the loss on the specific tranche; in this case the general order of payments in the interest waterfall is consecutively:

- interest on senior notes;
- replenishment of senior notes PDL;
- interest on mezzanine notes;
- replenishment of mezzanine notes PDL;
- (same for the junior and equity notes)
- replenishment of reserve fund;
- deferred purchase price to issuer.

2.5 Other common features

In this section we describe some other common features of RMBS transactions, namely subclasses, liquidity facilities, substitution and replenishment of the mortgage pool and redemption of the notes.

Within a tranche sometimes subclasses are indicated, where interest payments and default losses are equally distributed over the subclasses. However, repayment of principal is done sequentially, resulting in longer expected maturities for lower subclasses. Subclasses are common in the senior tranche.

The liquidity facility manages a timing mismatch between payments received from the mortgage pool and payments to be made to the noteholders. The SPV can temporarily draw money from the facility to bridge the timing mismatch. To ensure that a liquidity facility is not transformed to a credit enhancement tool, all amounts drawn from this facility are repaid to the liquidity provider at the top of the interest waterfall.

Two processes resulting in adding new mortgages to the underlying mortgage pool are substitution and replenishment. Substitution relates to substituting a mortgage which no longer meets the requirements on the mortgage pool set in the prospectus. This could for example happen if a borrower takes out a second mortgage on the same property. Some RMBS transactions specify an initial replenishment period in which no redemption of the outstanding notes occurs, instead prepaid mortgages are replaced by new mortgages.

Finally, the issuer has in general some freedom in determining the date at which the transaction is called and the remaining outstanding notes are redeemed in full. If the issuer decides not to redeem at the first optional redemption date (FORD), a step-up margin will have to be paid out to the noteholders on top of the interest payments at each payment date following the FORD. Every consecutive payment date until the final maturity of the RMBS is an optional redemption date. Besides redemption after the FORD, issuers can frequently also exercise a clean-up call option, which is the option to redeem all notes before the FORD when only a small portion, for example 10%, of the initial balance is still outstanding.

Chapter 3

Framework of RMBS valuation tool

In this chapter we describe the framework of the RMBS pricing tool developed in this project. We will follow the approach outlined by McDonald et al, (2010), whom developed a pricing model for mortgages in the UK mortgage market. Although our purpose is not to price mortgages directly, the value of an RMBS transaction heavily depends on the underlying mortgages. The model by McDonald et al. estimates the probability of default on a month-to-month basis at customer level, and applies this information to conduct a Monte Carlo simulation on the cash flows from a mortgage. We will extend the mentioned model by explicitly incorporating the competing events of mortgage termination by default and early repayment; details are supplied in the next chapters. The first section discusses in general the modelling of cash flows and the second section describes in greater detail the steps in the simulation process.

3.1 Cash flow modelling

The goal of this research is to develop a valuation and risk management tool for notes of an RMBS transaction that NIBC holds in its portfolio.

Subsequently the valuation of a note can be compared to the price offered in the market. We define the value of a note as the net present value (NPV) of all cash flows to the relevant tranche divided by the number of notes in the tranche. Our interest is solely in the NPV at the time of issue of the financial product, which we will define as time t_0. In general the NPV of a financial product is the sum over all discounted cash flows:

$$NPV_{t_0} = \sum_{t=1}^{m} CF_t \cdot df_t \,, \tag{3.1}$$

where the summation over the payment dates t extends over the interest and notional cash flows (CF) of the note and df is the discount factor. To make this more precise, let us define, in analogy with Burkhard and De Giorgi (2004), by $\mathbf{W} = (\mathbf{W}_t)_{t \geq t_0} = \{(d_i, B_i, V_i, I_i, L_i), i = 1, \ldots, n\}$ a portfolio of n mortgages outstanding during some period after time t_0. The process $\mathbf{W}$ is defined on a complete probability space $\{\Omega, (F_t)_{t \geq 0}, P\}$, with $(F_t)_{t \geq 0}$ a right-continuous filtration. For mortgage i, d_i denotes the time of origination, $B_i = (B_{i,t})_{t \geq d_i}$ is a process giving the outstanding balance at time t, $V_i = (V_{i,t})_{t \geq d_i}$ is a stochastic process representing the house value at time t, $I_i = (I_{i,t})_{t \geq d_i}$ is the process (stochastic or deterministic) describing the contract rate due on mortgage i and finally, $L_i = (L_{i,t})_{t \geq d_i}$ stands for any further information available on borrower i, such as his income and the location of the property. We assume that a mortgage portfolio is completely characterized by $\mathbf{W}$. Also define the stochastic interest rate process $r = \{r_t | \, t \in [0, T]\}$ on the same probability space. The cash flows to the noteholders depend on defaults and early repayments in the underlying mortgage pool which is described at each time t by the stochastic process $\mathbf{W}_t$. The actual value of the cash flows at time t_0 to the investors is determined by the discount factor, which is a function of r. Hence, we can write the expectation of the NPV of the cash flows to a tranche as

$$E[NPV_{t_0}(\mathbf{W}, r)] = E\left[\sum_{t=1}^{m} CF(\mathbf{W}_t) \cdot df(r_t)\right] . \tag{3.2}$$

The function $CF(\cdot)$ is not linear or continuous in $\mathbf{W}$, therefore it is very hard

to solve this expectation directly. Consequently we will revert to a (Monte Carlo) simulation. To this end we generate a large number of independent realisations W^i and r^i, $i = 1 \ldots, N$ of the respective random process $\mathbf{W}$ and r and calculate the sample average

$$\frac{1}{N}\sum_{i=1}^{N} NPV_{t_0}(W^i, r^i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{m} CF(W_t^i)\cdot df(r_t^i)\,. \tag{3.3}$$

We can safely assume that $E[NPV_{t_0}(\mathbf{W}, r)] < \infty$ and therefore we can conclude from the strong law of large numbers by Kolmogorov that with probability 1 it holds that

$$\frac{1}{N}\sum_{i=1}^{N} NPV_{t_0}(W^i, r^i) \to E[NPV_{t_0}(\mathbf{W}, r)] \text{ as } N \to \infty\,. \tag{3.4}$$

See for more details on Monte Carlo simulation and the corresponding properties and techniques, Krystul (2006) or Caflisch (1998). More important from a risk perspective, is calculating a probability distribution of the NPV of a note at t_0 such that the uncertainty in the value of an RMBS transaction can be quantified. Having run the model for N iterations one is left with N potential cash flow forecasts for the loan portfolio. From the allocation of these cash flows we can calculate the distribution of NPV_{t_0} of the notes in the RMBS transaction. Note that this approach is far more comprehensive than the approach used by rating agencies. Rating agencies simply use a rating scale to express the risk in a bond from a loss perspective. Thus, a AAA rating estimates the risk of a loss (i.e. missed payment) as less then 0.01%. However, the NPV of a note also depends upon the interest rates used in discounting and can therefore change without a missed payment. The approach we use gives the complete distribution function and thereby distincts itself from the rating agencies.

3.2 Simulation process

In the simulation process outlined in the previous section we need to predict the cash flows from the underlying mortgage pool. To this end we will predict

the state of the mortgage pool at time $t+1$ based on the state at time t and roll this forward from issue date to maturity. This process is outlined in figure 3.1 and explained step by step below.

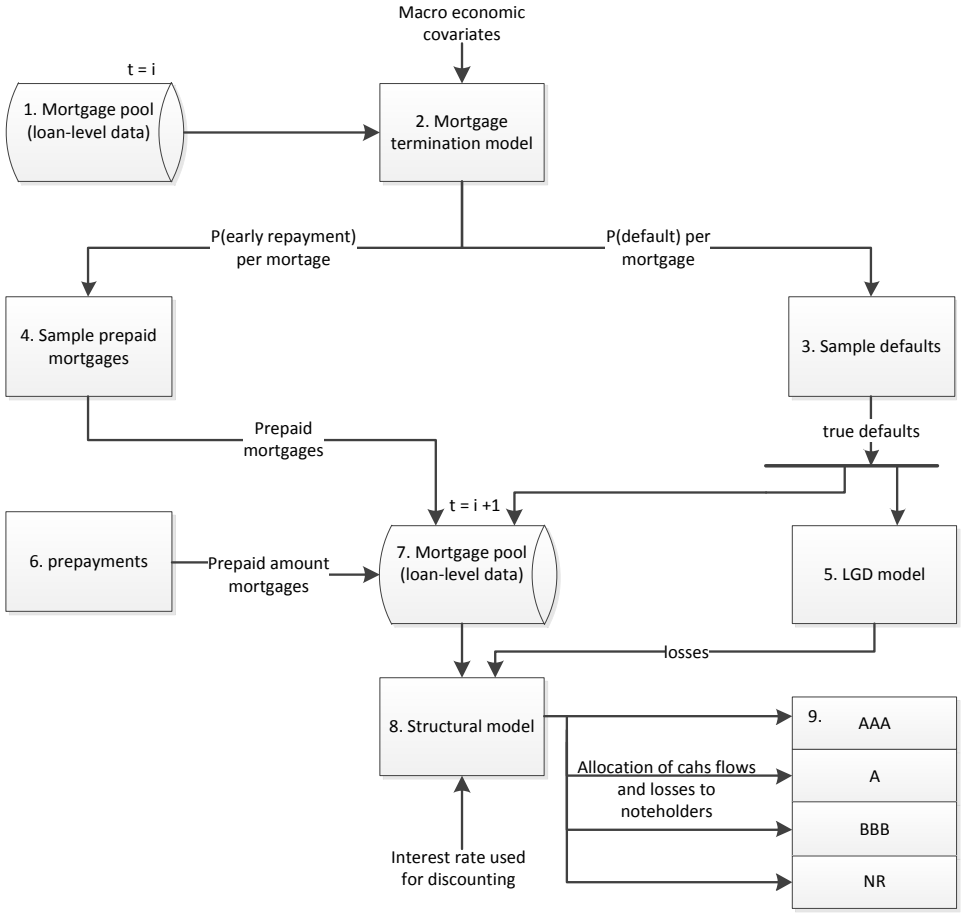


Figure 3.1: Outline of simulation process

1. We will describe the mortgage pool by data for each individual mortgage. Calculation time will of course increase significantly compared to using aggregated data, but since this process is not part of daily business, it is not really problematic.

2. The loan-level data of a mortgage is then used as input for the mortgage termination model, which will calculate the probability of default as

well as the probability of early repayment for each individual mortgage. The details of this model are outlined in the chapters 5 and 6.

3. From the probabilities of default we can sample which mortgages will actually default.

4. In the same way we can sample the mortgages that will be prepaid, where prepaid mortgages are defined as mortgages that are fully repaid before maturity.

5. The sampled defaulted mortgages are used as input in the loss-given-default (LGD) model. From this model we obtain the loss for each mortgage; this is further discussed in section 4.2 and section 8.3. The result of this step will give us the loss on the portfolio.

6. We can sample for each mortgage an amount prepaid, which in contrast to early repayment is only a partial repayment of the mortgage debt. As will be discussed in section 4.1 we will, in this project, assume these prepayments to be zero.

7. Based on the sampled early repayments and defaults, it is now possible to describe the mortgage pool at the next payment date.

8. The description of the mortgage pool and the losses incurred are then used as input for the structural model. This model describes the transaction specifics, such as the triggers in the waterfall structure, the size of the tranches and the return on the notes.

9. Finally the losses and cash flows from the underlying mortgage pool can be allocated to the different tranches.

One simulation consists of the mentioned process rolled forward to maturity, such that we obtain the NPV at the issue date of a note in each tranche under a specific realisation of the stochastic processes. The simulation is run for N iterations, which the user may vary according to computational resources

available. Finally we obtain a probability distribution for the NPV_{t_0} for each tranche.

Note that the developed model is limited to RMBS transactions based on Dutch mortgages. It is not realistic to assume that a model calibrated on Dutch mortgages is also valid for other markets. The main reason for this is that the parameters of the model will be different due to other characteristics of the underlying market. For example in the Netherlands a borrower can deduct the interest payments on his mortgage loan from his taxable income. This effect is not explicitly modelled, but it does keeps prepayment lower than it would be without this tax regulation. In other words, the interpretation of the parameters is restricted to actual study conditions and these differ for other countries too much from those in the Netherlands.

Chapter 4

Modelling mortgage cash flows

Valuation of RMBS transactions requires modelling the size and timing of cash flows from the underlying mortgage pool. To this end we need a model for the probability that a borrower will default. The actual loss when a borrower defaults, called loss-given-default (LGD), depends primarily on the amount still outstanding, the value of the underlying property and the probability of recovery. Recovery takes place when a defaulted borrower starts to pay his debt again, such that the default does not result in a loss to the issuer.
Since the value of a note also depends on the timing of cash flows, we also need to model early repayment of mortgage loans. In this chapter we give an outline of the existing literature on default and early repayment models for mortgages and we also briefly discuss LGD models.

4.1 Termination of mortgage loans by default or early repayment

A mortgage may be terminated before the legal maturity for two distinct reasons, either the mortgage is prepaid or the borrower defaults on his payment obligations. The most important reasons for fully prepaying a mortgage are house sale and refinancing the mortgage loan by taking out a new mortgage

against a lower interest rate. A borrower can redeem part of his mortgage if he has excess money and wants to lower his debt. As Alink (2002) points out these kinds of extra prepayments, although they are quite common, only account for around 5% of the cash flows resulting from prepayment. For this reason we will, in this project, assume that these kind of prepayments are nonexisting. We will concentrate on modeling the probability that a borrower fully prepays his mortgage and refer to this as early repayment in contrast to partial redemption of a mortgage which we will refer to as prepayment.
Default is another important feature of mortgage loans. When payments on a mortgage loan are first missed, the lender considers that the borrower is only temporarily delaying payment with the intention of renewing payment in the future, at which point the borrower is said to be in delinquency (Quercia and Stegman, 1992). It is the lender who decides when default has happened. We will use the definition for default from Basel II, which states that a borrower is in default if he is more than 90 days in arrears, i.e. the borrower has not made any interest or principal payments on his mortgage obligation for more than 3 months.
Essentially, there are two alternative views of residential mortgage default (Jackson and Kasserman, 1980), which are closely related to the two different ways of analysing early repayments: the equity theory and the ability-to-pay theory. We will discuss both these theories and discuss which one is most appropriate for our purpose.

4.1.1 Equity theory

The equity theory of default (also called option-theoretic view), assumes that borrowers will behave economically. Any mortgage contract contains two options: the prepayment option and the default option. A rational borrower will base his default decision on a comparison of the financial cost and returns involved in continuing or terminating mortgage payments. This view explicitly models defaulting on the mortgage as a put option on the underlying asset, where borrowers are hypothesized to exercise the option when

their equity position becomes negative. In this case the borrower sells back his house to the lender in exchange for eliminating the mortgage obligation. Early repayment is considered a call option, i.e. an option to buy back the mortgage at par. The ancestor of all option based models of default is the model by Merton (1974). Early contributions based on this idea are by Foster and Van Order (1985), Epperson et al. (1985) and Hendershott and van Order (1987). While these assumptions might seem appropriate for commercial borrowers, they are not that realistic when considering residential borrowers. A private individual's purpose is to finance his property with the mortgage and therefore his behaviour will not always be rational in the economic theory sense. An even more important shortcoming of the equity theory arises when we consider the legal aspects of a mortgage contract. The majority of these models were developed in an attempt to describe the credit risk of the mortgage market in the United States. While in the U.S. the originator of the loan only has rights on the property in case of default, this is different in Europe where mortgage lenders have full recourse to the borrower. That is, if a borrower defaults on his mortgage and the proceeds from the foreclosure do not cover the outstanding principal amount, the lender may chase the borrower for the shortfall on the market value of the property and the outstanding mortgage amount. For example, in the Netherlands a lender is able to seize a portion of the borrower's earnings from his employer in case the borrower defaults (Dutch MBS prospectus, 2005). Note that although in the Netherlands the law of remission of debt (in Dutch: wet schuldsanering) can restrict the actual recourse on the lender we will not account for this in our model. Also, among others Kau and Slawson (2002) report that borrowers do not exercise early repayment options optimally and that most practitioners do not believe in optimal prepayments. In the Netherlands it is common practice that borrowers pay a prepayment penalty when they repay their mortgage before maturity on another date than an interest reset date. Therefore it is not suitable for our purpose to model default or early repayment as an option on the value of the property.

4.1.2 Ability-to-pay-theory

The ability-to-pay theory of default states that borrowers refrain from loan default as long as income flows and cash reserves are sufficient to meet the periodic payments. Models based on this view are therefore much less economical and are based on empirical research. Within the ability-to-pay theory there is a wide variety of models, but popular ways of modelling defaults are binary choice models and survival models.

Binary choice models use a dependent variable which takes the value one if a certain event happens and zero otherwise. It models binomially distributed data of the form $Y_i \sim B(n_i, p_i)$ for $i = 1, 2, \ldots, m$, where the number n_i of Bernoulli trials are known and the probabilities of success p_i are unknown. In our case we would define 'success' as the event of a default or an early repayment. Two common variants of binary choice models are the probit model and the logit model, where respectively the inverse cumulative distribution function and the logit function $\left(\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)\right)$ are assumed to be linearly related to a set of predictors. The probit model is among others used by Webb (1982) to differentiate probability of default among different mortgage instruments. Campbell and Dietrich (1983) apply the logit model to residential mortgages in the U.S. and Wong et al. (2004) apply it to residential mortgages in Hong Kong.

Survival models deal with the distribution of survival times. Although there exists some well-known methods to estimate the unconditional survival distribution, more interesting models relate the time that passes before a certain event occurs to one or more explanatory variables. In our case the event of interest would be the termination of a mortgage, either by early repayment or by default. Since both causes of termination have their own specific effect on the value of the mortgage to the lender, we want to be able to estimate these probabilities separately. In this case we speak of a competing risk setting, where the occurrence of default (early repayment) prevents the occurrence of early repayment (default). Survival models explicitly incorporate the altering probability of occurrence of an event with time. This is essential in

a mortgage setting, where presumably both the probability of default and the probability of early repayment are not constant over time. Although this could also be achieved in, for example, a logit model by incorporating age as an explanatory variable, this type of model is less informative and intuitively understood. Therefore we will in this project apply the survival approach to model the probability of default as well as the probability of early repayment; details of this approach will be supplied in the next two chapters.

4.2 Loss-Given-Default

As mentioned before, a mortgage is considered to be in default when no interest or principal payments have been made for more than three months. The process of foreclosure can than be started by the originator of the loan. Foreclosure is a legal process, which targets to sell the property so that the proceeds can be used to meet the contractual obligations of the mortgage contract. This process can take between a few months to over a year depending on the jurisdictions of a country. Loss-given-default (LGD) is the incurred loss when default happens and includes the unpaid balance, accrued interest, legal foreclosure expenses, property maintenance expenses and sales costs. This definition resembles the Basel II definition. LGD is equal to exposure at default (EAD) $\cdot$ (1 $-$ the recovery rate). This recovery rate is in literature most often modelled by an U-shaped beta distribution. An extensive research on the LGD is outside the scope of this research; we will therefore approach this issue from a more practical point of view and return to this point in section 8.3.

Chapter 5

Survival analysis

Survival analysis is the area of statistics that deals with the analysis of life time data and has its origin in medical and reliability studies concerned with the failure time of machines and devices. As shown by Banasik et al. (1999) and McDonald et al. (2010) it is also applicable to estimate the time to both default and early repayment of mortgages. The major strength of survival analysis is the ability to incorporate censored data; observations for which the event of interest does not take place in the sample period. The best known survival model is the Cox proportional hazards model, which we will apply in this project. To be able to incorporate the competing risk of terminating a mortgage by either default or early prepayment, where occurrence of one event rules out the possibility of occurrence of the second event, we need to adapt the Cox model. This chapter will start by explaining the basics of a survival model. Then we will discuss the Cox proportional hazards model in the absence of competing risks and finally we will discuss adaptations to the Cox model in a competing risks setting.

5.1 Definition and formulas

The general terminology in survival analysis speaks of subjects, which are in our cases mortgages, and an event or a failure which is for our model

the termination of a mortgage by either default or early repayment. In this chapter and the next we will use the survival analysis formulation when explaining a general concept and the formulation in terms of mortgages when applying the model to our problem.

In this research, we will apply survival analysis to model the probability of default or early repayment of a mortgage. Both default and early repayment are causes of terminating a mortgage. By using survival analysis we can relate the lifetime of a mortgage to certain characteristics of the loan. The probability density function of the survival time of a mortgage, i.e. the time to termination of a mortgage, with certain age and characteristics gives a direct way to find the month-to-month probabilities of termination. We will derive the probability density function in this section.

Let us consider a random time τ defined on a probability space $(\Omega, \mathcal{F}, P)$, i.e. $\tau : \Omega \to (0, \infty)$ is a positive continuous $\mathcal{F}$-measurable random variable. Note that τ is a stopping time. We can interpret τ as the time to termination of a mortgage. We denote by $f(t)$ the probability density function of τ, i.e.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t)}{\Delta t} \tag{5.1}$$

and by

$$F(t) = P(\tau \leq t) = \int_0^t f(u)\, du \tag{5.2}$$

the cumulative distribution function of τ. We have assumed here that $F(t)$ is absolutely continuous. The survival function measures the probability of no occurrence of the event till time t, i.e.

$$S(t) = 1 - F(t) = P(\tau > t). \tag{5.3}$$

We assume that $F(0) = P(\tau = 0) = 0$ and that $S(t) > 0$ for all $t < \infty$.

The most important function of survival analysis is the hazard rate.

Definition 5.1. *Hazard rate can be interpreted as the time-specific failure rate and can formally be expressed as a ratio of the conditional probability for*

the event to occur within an infinitely small interval over the time interval, as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t | \tau \geq t)}{\Delta t}. \tag{5.4}$$

By this definition, the hazard rate $\lambda(t)$ measures the rate of change at time t. Note that hazard rates can exceed the value one. The cumulative hazard function is the integral of the hazard rate from time 0 to time t,

$$\Lambda(t) = \int_0^t \lambda(u)\, du. \tag{5.5}$$

One can express $\lambda(t)$ as a function of $f(t), F(t)$ and $S(t)$ as follows:

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t | \tau \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t)/\Delta t}{P(t \leq \tau)} \\ &= \frac{f(t)}{S(t)} = \frac{-\frac{d}{dt} S(t)}{S(t)} = -\frac{d}{dt} \log S(t), \end{aligned} \tag{5.6}$$

where the last equality follows from the chain rule. So we also have

$$S(t) = e^{-\Lambda(t)}. \tag{5.7}$$

If we want to determine the survival function without accounting for characteristics of a specific mortgage, we can estimate it by the well-known Kaplan-Meier estimator (Kaplan and Meier, 1958). With this estimator every mortgage has the same probability of termination during it's lifetime, without distinguishing between mortgages based on characteristics other than age. The Kaplan-Meier method starts by sorting the event times in an ascending order; we denote the rank-ordered failure times $\tau_{(1)} < \tau_{(2)} < .. < \tau_{(m)}$. Now we will give a definition for *risk set*, since we will encounter this term more often in this chapter and the next.

Definition 5.2. *The risk set at time t is a set of indices of all subjects (mortgages) that are 'at risk' of failing (defaulting or early repaying) at time t. Thus the risk set contains all subjects which did not fail before time t.*

For the Kaplan-Meier estimator let the risk set at time $\tau_{(i)}$ be denoted by n_i, so n_i are all mortgages still performing at time $\tau_{(i)}$. We denote the observed number of failures at time $\tau_{(i)}$ by d_i. The Kaplan-Meier estimator of the survival function at time t is

$$\widehat{S}(t) = \prod_{\tau_{(i)} \leq t} 1 - \frac{d_i}{n_i}\,. \tag{5.8}$$

5.2 Censoring

Before going into more details on survival analysis, we first have to describe censoring. Censoring refers to a situation where exact event times are known only for a portion of the study subjects (Guo, 2010). The ability of survival techniques to cope with censored observations gives them an important advantage over other statistical techniques. It is nearly impossible to analyse the duration of a mortgage without including censored ones. Their absence would necessitate at least 30 years of historical data, which is the legal maturity of a typical Dutch mortgage contract. To describe what a censored observation is, it is easiest to describe first an uncensored observation.

Definition 5.3. *An uncensored time-observation of the life-time of a mortgage, starts at the issue date (t=0) and ends when the mortgage is terminated by default or early repayment.*

So, for an uncensored observation of a mortgage all covariates are known over its lifetime and it is terminated at a known time point by either default or early repayment. An observation of a mortgage which never defaults and pays of the loan at maturity is by this definition not an uncensored observation, even though the entire lifespan of the mortgage is observed.
The most common type of censoring is when the subject has not experienced an event at the end of the observation period. This type of censoring is called right-censoring. Although, we do not know for a right-censored mortgage observation if the mortgage will ever default or early repay, we obtain

the information that the mortgage has survived at least until the time of censoring. Survival analysis techniques use this information in fitting a model; this wil be discussed in section 6.1.

Other types of censoring are left-censoring and left-truncation (or delayed entry). Left-censoring occurs when an event is known to have happened before the sample period starts, however the exact event time is unknown. We speak of left-truncation when $t = 0$ is preliminary to the start of the observation period. In this case it may happen that subjects with a lifetime less than some threshold are not observed at all. In a so called *delayed entry* or (left-truncated) study, subjects are not observed until they have reached a certain age. The type of censoring in which observations are both left-truncated and right censored is called interval-censoring. This research will examine three types of censored observations in addition to the uncensored observations, namely; left-truncated, right-censored and interval-censored observations. In figure 5.1 the different types of observations that we encounter are displayed graphically. The observation period starts July 2004 and ends December

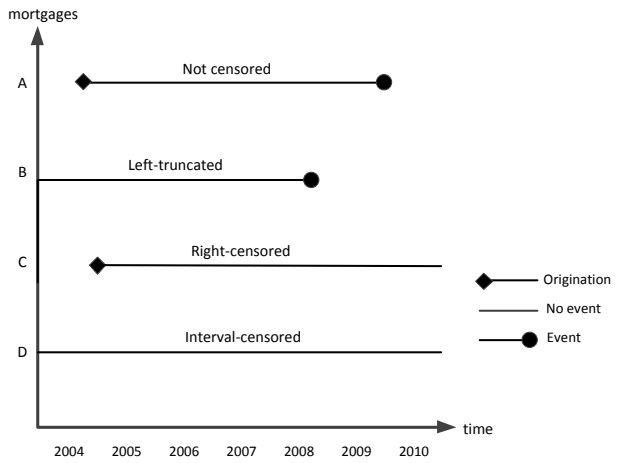


Figure 5.1: Different types of censoring

2010, while a portion of the mortgages in the sample are issued before July 2004. This fact makes our study to a typical delayed entry study; we will discuss in section 6.5 specific issues arising in such a study and how it can

be dealt with to prevent biasing the estimated probabilities. A description of the data is given in section 7.1.
We have to discuss the different reasons of occurrence of right-censored data in more detail, as it will be of importance later on. Right-censored event times can be categorized as (Putter et al, 2007):

- End of study: the event has not yet happened at the end of the sample period. This is also called administrative censoring.
- Loss to follow-up: the subject left the study due to other reasons. The event may have happened but this information is unknown.
- Competing risk: another event has occurred, which prevents occurrence of the event of interest.

If the reason of censoring is "'end of study"' then we can, in general, safely assume that the censoring mechanism is independent of the event time. In the other two situations we should be more careful. Right-censored data plays an important role in survival models, but when the censoring mechanism can be assumed to be independent of the event time, it can be dealt with fairly easily. We will return to this point later.

5.3 Cox proportional hazards model

In this section we will present a way to model the hazard rate. The Cox proportional hazards model is a well known survival model mostly applied in medical science to model the relationship between the survival of a patient and one or more explanatory variables (called covariates). We will in this section explain the basic model by first assuming that the censoring mechanism is independent of the event times. In the next section we will discuss the competing risks setting and thereby relax this assumption.
The Cox proportional hazards model (sometimes abbreviated to Cox model) was first proposed by Cox (1972) to extend the results of Kaplan and Meier

(1958) by incorporating covariates in the analysis of failure times. The name of the model comes from the feature that the ratio of the hazard rates of two subjects is constant over time.

Definition 5.4. *The Cox proportional hazards model can be expressed as:*

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp(\boldsymbol{\beta}^T \mathbf{X}), \tag{5.9}$$

were $\lambda(t|\mathbf{X})$ *is the hazard rate conditional on a vector of covariates* $\mathbf{X} = (X_1, .., X_p)$ *and* $\boldsymbol{\beta} = (\beta_1, .., \beta_p)$ *gives the influence of these covariates on the hazard rate.* $\lambda_0(t)$ *is the baseline hazard function and can be thought of as the hazard rate for an individual whose covariates all have value zero.*

The proportional hazards model is non parametric in the sense that it involves an unspecified function in the form of an arbitrary baseline. In this model a unit increase in a covariate has a multiplicative effect with respect to the hazard rate. To be exact, if X_j increases by one unit the hazard rate is multiplied by a factor e^{β_j}.

5.4 Time-varying covariates

Until now we have assumed that the values of all covariates were determined at the starting point of the study and that these values did not change over the sample period. It is also possible to explicitly account for changes to one or more covariates during the sample period by making use of time-varying (or time-dependent) covariates. The basic idea behind time-varying covariates requires thinking in terms of a 'counting process' setup; for details on the counting process formulation of the Cox model see Andersen and Gill (1982). In this setup, each record (line of data) gives the value of covariates that are constant between two time points, and whether the event of interest took place by the ending time point or not. In our model we intend to incorporate covariates that might change every month. Consequently, our data will consist of a number of records equal to the number of observed months,

i.e. outstanding months during the sample period, for each mortgage.
In analogy with Hosmer et al. (2008) we can classify time-varying covariates as being either internal or external. An internal time-varying covariate is subject specific and therefore requires that the subject is under direct observation. In contrast, an external time-varying covariate is typically a study or environmental factor which applies to all subjects in the sample. This type of covariate does not require periodic observations of the subject. Both types of time-varying covariates might play a role in our model. An example of an internal time-varying covariate is the outstanding balance of the mortgage loan and an example of an external time-varying covariate is the unemployment rate. We can write the Cox model with time-varying covariates as

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp\left(\boldsymbol{\beta}_1^T \mathbf{X}_1 + \boldsymbol{\beta}_2^T \mathbf{X}_2(t)\right), \tag{5.10}$$

where $\mathbf{X}_1$ is the vector with constant covariates and $\mathbf{X}_2(t)$ consists of the time-varying covariates.
From a conceptual point of view the model becomes much more complicated by introducing time-varying covariates. Specifically, it causes an inability to give individualized predictions of the estimated event time when the future values of time-varying covariates are unknown. We can deal with this by making strict assumptions on the progress of the covariates or simulate possible paths. A discussion with elaborate examples on the issue of time-varying covariates can be found in Fisher and Lin (1999).

5.5 Competing risk models

Another useful generalization of survival models is the concept of competing risks. This assumes that a subject can fail due to more than one reason, but only the first to occur can be observed. In our case there are two competing risks, namely default and early repayment, which are both of interest to us. In this case we extend the setting of no competing risks by supposing that the n subjects give rise to the data $(\tau_i, \delta_i, \epsilon_i, \mathbf{X}_i), i = 1, \ldots, n$ where

again τ_i is the observed survival time, δ_i is the censoring indicator ($\delta_i = 0$ if the i-th subject is right-censored and 1 if any of the m competing events happened), ϵ_i is the failure type ($\epsilon_i = 1, .., m$) and $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,p})$ is the covariate vector of the i-th subject. For our model, n is the number of mortgage observations, $m = 2$ and $\epsilon_i = \text{default}, \text{early repayment}$.
Note that the Kaplan-Meier estimator in equation (5.8), used to estimate the overall survival distribution, treats the competing risks as censored. The probability of failure due to a specific cause is overestimated if the competing risks are not independent, which is due to the fact that the independent censoring assumption is not fulfilled, see Klein and Bajorunaite (2004), Tai et al. (2001) and Satagopan (2004) for details.

5.5.1 Overview of competing risk literature

Competing risk models can be classified to belong to one of two classes. The first one assumes that there are m hypothetical failure times, whereas the second is related to the joint distribution of time τ and cause j of failure.
The first approach, often called latent failure time approach, views competing risk models as a multivariate failure time model, where each subject is assumed to have a potential failure time for each failure type. The earliest of these failures is actually observed and the others are latent, i.e. $\tau_i = \min(\tau_{i,1}, \tau_{i,2}, ..., \tau_{i,m})$ or for our model $\tau_i = \min(\tau_{i,\text{default}}, \tau_{i,\text{early repayment}})$. Although this view gives a nice physical interpretation to latent failure times as potential failure times, it also has some serious drawbacks. The latent failure time approach involves the very strong assumption that the time of failure from cause j under one set of study conditions in which all m causes are present is exactly the same as under an altered set of conditions in which all causes except the j-th have been removed; this is only the case if the competing risks are independent of each other. According to Prentice et al. (1978) it has been long recognized that the elimination of certain failure types may well alter the risks of other types of failure. Since it is undesirable to assume independence of default and early repayment, we will not discuss

this approach further.
The second approach to competing risks is more recent and it deals with the existence of failure times $\tau_{i,1}, \tau_{i,2}, .., \tau_{i,m}$ on each subject i under the actual study conditions. The random variable $\tau_{i,j}$ is the observed time of failure of subject i due to cause j, and there is no physical interpretation attached to the unobserved $\tau_{i,k}$'s. Concretely, this means that when a mortgage defaults, no information on the potential early repayment time for this mortgage is obtained. One of the earliest attempts to account for a form of informative censoring was presented by Kimball (1969), where subjects that failed from competing risks will fail from the event of interest with probabilities related to those obtained before the competing risk was eliminated. This approach is somewhat arbitrary and we will therefore not further discuss it. Two methods of the second approach often applied in the context of possible dependent competing risks, are the cause-specific hazard rate (Kalbfleisch and Prentice, 1980 and Prentice et al, 1978) and the subdistribution hazard rate (Fine and Gray, 1999). We will discuss both methods methods in more detail now.

5.5.2 Cause-specific hazard rate

The classical approach of Kalbfleisch and Prentice defines the cause specific hazard rate in the presence of competing risks as an evident extension to the ordinary hazard rate of definition 5.1.

Definition 5.5. *The cause-specific hazard rate of cause j in the presence of competing risks is defined as*

$$\lambda_j(t|\mathbf{X}) = \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t, \epsilon = j | \tau \geq t, \mathbf{X})}{\Delta t} \text{ for } j = 1, \ldots, m \quad (5.11)$$

where $\mathbf{X}$ is the regression vector.

For our purpose we define τ as the time to termination of a mortgage and $j = 1, 2$ refers to respectively default and early repayment of a mortgage. Observations of subjects for which $\epsilon \neq j$ are treated as censored at the time of termination in the same way actual right-censored observations

are treated. We may, similar to the ordinary Cox model, define the cumulative cause-specific hazard function by $\Lambda_j(t) = \int_0^t \lambda_j(s)\,ds$ and define $S_j(t) = \exp(-\Lambda_j(t))$. Note that, although $S_j(t)$ can be estimated it should not be interpreted as a marginal survival function; it only has this interpretation if the competing event times and the censoring times are independent. In that case, the marginal distribution describes the event time distribution in the situation that the competing events do not happen. We can also define $S(t) = \exp(-\sum_{j=1}^{m} \Lambda_j(t))$, which does have a clear interpretation as the probability of not having failed at time t from any cause.
Based on the Cox model, we assume that the cause-specific hazard rate has the following form:

$$\lambda_j(t|\mathbf{X}) = \lambda_{0,j}(t) \cdot \exp(\boldsymbol{\beta}_j^T \mathbf{X}), \tag{5.12}$$

where $\lambda_{0,j}(t)$ is the cause-specific baseline function of cause j, and the vector $\boldsymbol{\beta}_j$ represents the covariate effects of cause j. Note that for our model $j = 1, 2$ refers to respectively default and early repayment of a mortgage. The interpretation of the effects of the covariates is restricted to actual study conditions and there is no implication that the estimates would remain the same under a new set of conditions. This implies, for example, that if the deductibility of mortgage loan interest from taxable income, which is currently a highly popular income tax policy in the Netherlands, would be restricted by law, the obtained model for the probability of default is no longer valid. Due to a lower available income the probability of default will presumably increase.

Another implication worth mentioning is the impossibility to directly predict the effect of the covariates on the cumulative incidence function, see definition 5.6.

Definition 5.6. *The cumulative incidence function (CIF) of cause j, is defined as the probability of failing from cause j before time t; it can be expressed as*

$$F_j(t) = P\left(\tau \leq t, \epsilon = j\right) . \tag{5.13}$$

The CIF can be expressed in terms of the cause-specific hazards function as:

$$\begin{aligned} F_j(t) &= P\left(\tau \leq t, \epsilon = j\right) \\ &= \int_0^t \lambda_j(s) P(\tau \geq s)\, ds \\ &= \int_0^t \lambda_j(s)\, S(s)\, ds \\ &= \int_0^t \lambda_j(s) \exp\left(- \int_0^s \sum_{i=1}^{m} \lambda_i(u)\, du \right) ds. \end{aligned} \tag{5.14}$$

Note that we left out, for ease of notation, the conditioning on the regression vector $\mathbf{X}$. From equation (5.14) it follows that the cumulative incidence function for cause j does not only depend on the hazard rate of cause j, but also on the hazard rates of all other causes. In other words, the probability of default for a mortgage does not only depend on the hazard rate of default, but also on the hazard rate of early repayment and vice versa.

5.5.3 Subdistribution hazard rate

In order to avoid the highly non-linear effect of covariates on the cumulative incidence function in the approach by Kalbfleish and Prentice, Fine and Gray introduced a way to directly regress on the CIF. Gray (1988) defined the subdistribution hazard rate in a bit trickier way than the cause-specific hazard rate.

Definition 5.7. *The subdistribution hazard rate for cause j in the presence of competing risk is defined as*

$$\alpha_j(t|\mathbf{X}) = \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t, \epsilon = j | \tau \geq t \cup (\tau \leq t \cap \epsilon \neq j), \mathbf{X})}{\Delta t} \tag{5.15}$$
$$\text{for } j = 1, \ldots, m\,.$$

where $\mathbf{X}$ is the regression vector.

Again τ is the time to termination of a mortgage by any of the competing risks. The difference between the cause-specific hazard rate and the subdistribution hazard rate is the definition of the risk set. In the first case the risk

set at time t exists only of those subjects that did not fail from any cause by time t. Whereas in the latter case the risk set at time t includes subjects that did not fail from any cause by time t and, in addition, the subjects that have previously failed from competing risks. This means that observations of subjects for which $\epsilon \neq j$ are not treated as censored as is the case for the cause-specific hazard rate approach. Clearly, the risk set associated with the subdistribution hazard rate α_j is unnatural, as in reality those subjects that have already failed due to another cause than $\epsilon = j$ prior to time t are not "at risk" at time t any more. We can define the random variable

$$\tau_j^* = \begin{cases} \tau & \text{if} \quad \epsilon = j \\ \infty & \text{if} \quad \epsilon \neq j \end{cases}$$

and write $\tau = \min(\tau_1^*, \tau_2^*, ..., \tau_m^*)$. One can think of α_j as the hazard rate for τ_j^*. The distribution function of the implied failure time τ_j^* can be written in terms of the CIF $F_j(t)$ as

$$\begin{cases} F_j(t) & \forall\, t < \infty \\ P(\tau_j^* = t) = P(\tau < t, \epsilon \neq j) = 1 - \lim\limits_{t \to \infty} F_j(t) & \text{for } t = \infty \end{cases} \tag{5.16}$$

Fine and Gray (1999) proposed a regression model based on Cox model by

$$\alpha_j(t|\mathbf{X}) = \alpha_{0,j}(t) \cdot \exp(\boldsymbol{\beta}_j^T \mathbf{X}). \tag{5.17}$$

The subdistribution hazard rate is by construction explicitly related to the cumulative incidence function by

$$\alpha_j(t) = \frac{-d \log(1 - F_j(t))}{dt}. \tag{5.18}$$

Note the close resemblance to the relationship in equation (5.6). As discussed in the previous subsection, the cause-specific hazard rate has a less clear relation to the cumulative incidence function and it involves the cause-specific hazard rates of failures from all other causes.

5.5.4 Choice of method

We intend to use a Cox proportional hazards model in a competing risks setting with time-varying covariates to estimate the probability of default and the probability of early repayment of a mortgage. From this section we can conclude that there are two candidate models for this purpose, namely the cause-specific hazard rate and the subdistribution hazard rate. In the cause-specific hazard rate approach, we can only calculate the cumulative incidence function of a cause by making use of all cause-specific hazard rates. This is not the case for the subdistribution hazard rate approach.
However, a big disadvantage of the model by Fine and Gray is that it can have some significant bias in the presence of two situations which are both applicable to our data, namely larger differences in the occurrence frequency of competing risks and time-varying covariates. The first case is evident from the definition of the risk set. Latouche et al. (2005) showed that the subdistribution hazard rate approach is not appropriate for estimating the effect of any time-varying covariate unless the entire path is observable. Since both default and early repayment terminate a mortgage, it would be hard to observe the entire path. Finally, the cause-specific hazard rate approach offers a much more intuitive interpretation of the risk set and the cause-specific hazard rate as the hazard rate of a cause in the presence of competing risks, where the subdistribution hazard rate has no physical interpretation. For these reasons we will implement the approach by Kalbfleisch and Prentice in this project.

5.6 Prediction

The purpose of applying survival analysis in our project is to be able to predict the probability of default and the probability of early repayment of a mortgage with specific age and characteristics. As discussed in the previous section we will follow the approach by Kalbfleish and Prentice to deal with the competing risks of terminating a mortgage by either default or early

repayment. We will be able to predict the cash flows originating from the mortgage pool, by going from one payment date of the RMBS to the next payment date and sample for all still outstanding mortgages whether they will default, early repay or remain performing. Our interest is therefore the monthly probability of default (or early repayment), conditional on survival up till this date, i.e.:

$$P(\tau = t, \epsilon = j | \tau > t-1) \qquad \text{for } j = 1, 2. \tag{5.19}$$

This equation can be rewritten as follows:

$$\begin{aligned} P(\tau = t, \epsilon = j | \tau > t-1) &= \frac{P\left((\tau = t, \epsilon = j) \cap (\tau > t-1)\right)}{P(\tau > t-1)} \\ &= \frac{P(\tau = t, \epsilon = j)}{P(\tau > t-1)} \\ &= \frac{P(\tau \leq t, \epsilon = j) - P(\tau \leq t-1, \epsilon = j)}{P(\tau > t-1)} \\ &= \frac{F_j(t) - F_j(t-1)}{S(t-1)}, \end{aligned} \tag{5.20}$$

where

$$\begin{aligned} S(t) &= \exp(-\Lambda_1(t) - \Lambda_2(t))\,, \\ \Lambda_j(t) &= \int_0^t \lambda_j(u)\, du\,, \end{aligned}$$

and $F_j(t)$ is defined as in equation (5.14). By filling in these quantities we can write

$$P(\tau = t, \epsilon = j | \tau > t-1) = \lambda_j(t) \exp\left(-\lambda_1(t) - \lambda_2(t)\right)\,. \tag{5.21}$$

The next chapter will describe the methods to estimate the model parameters, i.e. the vector $\boldsymbol{\beta}$ and the baseline $\lambda_0(t)$.

Chapter 6

Model estimation

In this chapter we describe the methods and formulas to estimate the model of chapter 5 for default and early repayment. For ease of notation we will in this chapter, whenever it does not cause ambiguity, leave out the subscript to denote the specific cause of failure and assume that covariates are constant over time. The first section describes the method to estimate the vector $\boldsymbol{\beta}$, the second section describes how to calculate the corresponding baseline using these estimates. The third section illustrates the methods by a simplified theoretical example. In the fourth section we discuss how to account for ties in the data. The last section discusses the design of our analysis and specifically the exact definition of the risk set in the context of our delayed entry study.

6.1 Parameter estimation

To estimate the influence of the covariates Cox (1972) applies a method called partial likelihood estimation, which discards the baseline function and only deals with the exponential part of the equation. We shall give a brief derivation of the maximum likelihood estimator for this method. Assume again that n study subjects give rise to the data $(\tau_i, \delta_i, \mathbf{X}_i)$, $i = 1, .., n$ where τ_i is the observed survival time, δ_i is the censoring indicator ($\delta_i = 0$ if the i-th

subject is right-censored and 1 if an event happened) and $\mathbf{X}_i$ is the covariate function for the i-th subject. Assume for the time being that there are no ties among the failure times and that we have observed $k \leq n$ mortgage terminations. First we sort the failure times in an ascending order, i.e. $\tau_{(1)} < \tau_{(2)} < .. < \tau_{(k)}$, as for the Kaplan-Meier estimator in equation (5.8). The likelihood function for subject i to have the event at time t is simply the hazard rate for subject i divided by the sum of the hazard rates for all subjects that are at risk of failing at time t, that is all mortgages outstanding at time t. So we can write the likelihood function for subject i as

$$\begin{aligned} PL_i &= \frac{\lambda_{(i)}(t|\mathbf{X})}{\lambda_{(i)}(t|\mathbf{X}) + \lambda_{(i+1)}(t|\mathbf{X}) + \cdots + \lambda_{(n)}(t|\mathbf{X})} \\ &= \frac{\lambda_{(i)}(t|\mathbf{X})}{\sum_{j=1}^{n} \mathbb{I}(j \geq i)\lambda_{(j)}(t|\mathbf{X})} \\ &= \frac{\lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}_{(i)})}{\sum_{j=1}^{n} \mathbb{I}(j \geq i)\lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}_j)} \\ &= \frac{\exp(\boldsymbol{\beta}^T\mathbf{X}_{(i)})}{\sum_{j=1}^{n} \mathbb{I}(j \geq i)\exp(\boldsymbol{\beta}^T\mathbf{X}_j)}\,, \end{aligned} \tag{6.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\mathbf{X}_{(i)}$ is the covariate vector corresponding to the i-th mortgage in ascending order. We can also define $\mathcal{R}(\tau_{(i)})$ as the risk set at time $\tau_{(i)}$, this means that $\mathcal{R}(\tau_{(i)})$ is a set of indices of all mortgages which are still performing at time $\tau_{(i)}$. Formally,

$$\mathcal{R}(\tau_{(i)}) = \{j = 1, .., n | \tau_j \geq \tau_{(i)}\},$$

so the mortgage i corresponding to failure time $\tau_{(i)}$ is itself also part of the risk set $\mathcal{R}(\tau_{(i)})$. We can now rewrite PL_i as

$$PL_i = \frac{\exp(\boldsymbol{\beta}^T\mathbf{X}_{(i)})}{\sum_{j \in \mathcal{R}(\tau_{(i)})} \exp(\boldsymbol{\beta}^T\mathbf{X}_j)}\,. \tag{6.2}$$

By multiplying the partial likelihood function for all n subjects we obtain the sample partial likelihood function, in which the likelihood function for

censored data is set to one:

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} PL_i$$
$$= \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_{(i)})}{\sum_{j \in \mathcal{R}(\tau_{(i)})} \exp(\boldsymbol{\beta}^T \mathbf{X}_j)} \,. \tag{6.3}$$

It is convention in statistics to take the logarithm of the likelihood function. Doing so, we seek to maximize

$$pl(\boldsymbol{\beta}) = \sum_{i:\delta_i=1} \left(\boldsymbol{\beta}^T \mathbf{X}_{(i)} - \ln \left[\sum_{j \in \mathcal{R}(\tau_{(i)})} \exp(\boldsymbol{\beta}^T \mathbf{X}_j) \right] \right) \,. \tag{6.4}$$

6.2 Baseline estimation

After we have obtained the estimates for $\boldsymbol{\beta}$ by maximizing the partial likelihood function of the previous section, we can plug in this estimate in the full likelihood function to obtain the corresponding baseline. Following this approach we find the best known estimator for the baseline, the so called Breslow estimator (Breslow, 1972). In this section we derive the formula of this estimator.

As in the previous section, assume for the time being that there are no ties among the failure times and let $\tau_{(1)} < \ldots < \tau_{(k)}$ denote the k distinct, ordered failure times in the sample set. Note that we have n observations, i.e. $k \leq n$. The full likelihood function for discrete measured failure times can be expressed as:

$$L(\boldsymbol{\beta}, \lambda_0(t)) = \prod_{i=1}^{n} \left[\lambda_0(\tau_i) \exp\left(\boldsymbol{\beta}^T \mathbf{X}_i\right) \right]^{\delta_i} \exp\left[- \sum_{u=0}^{\tau_i} \lambda_0(u) \exp(\boldsymbol{\beta}^T \mathbf{X}_i) \right] \,, \tag{6.5}$$

and the corresponding log likelihood function as:

$$l(\boldsymbol{\beta}, \lambda_0(t)) = \sum_{i=1}^{n} \delta_i \left[\ln(\lambda_0(\tau_i)) + \boldsymbol{\beta}^T \mathbf{X}_i \right] - \sum_{i=1}^{n} \lambda_0(\tau_i) \sum_{j \in R(\tau_i)} \exp(\boldsymbol{\beta}^T \mathbf{X}_j) \,. \tag{6.6}$$

The derivation of the full likelihood formula and the corresponding log likelihood formula can be found in appendix A. The vector $\boldsymbol{\beta}$ is replaced by the estimate $\hat{\boldsymbol{\beta}}$ and the only remaining unknown parameter in the likelihood function is the baseline function $\lambda_0(t)$. For discrete measured data it holds that if no event happened in the underlying data at time t, the model assumes that the probability of an event at time t is zero. Therefore we can conclude that $\hat{\lambda}_0(t) = 0$ for $t \notin \{\tau_{(1)}, \ldots, \tau_{(k)}\}$ and if all censored observations which occur in the interval between two consecutive events $(\tau_{(i)}, \tau_{(i+1)})$ are adjusted to have occurred at $\tau_{(i)}$, we can rewrite (6.6) as

$$l(\hat{\boldsymbol{\beta}}, \lambda_0(t)) = \sum_{i=1}^{k} \left[\ln(\lambda_0(\tau_{(i)})) + \hat{\boldsymbol{\beta}}^T \mathbf{X}_{(i)} \right] - \sum_{i=1}^{k} \lambda_0(\tau_{(i)}) \sum_{j \in R(\tau_{(i)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{X}_j) \,. \tag{6.7}$$

Maximizing (6.7) with respect to $\lambda_0(\tau_{(i)})$ gives the maximum likelihood estimate of $\lambda_0(\tau_{(i)})$ as

$$\hat{\lambda}_0(\tau_{(i)}) = \frac{1}{\sum_{j \in R(\tau_{(i)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{X}_j)} \,. \tag{6.8}$$

The baseline survival function estimate is given by

$$\widehat{S}_0(t) = \prod_{\tau_{(i)} < t} \left[1 - \frac{1}{\sum_{j \in R(\tau_{(i)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{X}_j)} \right] \,. \tag{6.9}$$

Note the close resemblance of $\widehat{S}_0(t)$ with the Kaplan-Meier estimator in equation (5.8) if the number of observed failures at one measurement time cannot exceed one.

6.3 An illustrating example

This section presents a simplified theoretical example to show the application of the approaches discussed in this chapter and the previous. We will assume that there are no time-varying covariates, no ties in the data and also there

are no delayed entries in the study. We assume that there is only one variable, the age of the borrower at time of issuance, which influences both the probability of default and the probability of early repayment of a mortgage. We assume furthermore that all mortgages are issued at the same time and that we are only interested in the first 40 months of the lifetime of a mortgage; our sample period is 40 months. In table 6.1 we give the details of our example and figure 6.1 displays this example graphically. A '1' in the column of default (early repayment) indicates that the borrower has defaulted (early repaid) at the end of the observation period. The column with time gives the observation time, either until the mortgage is terminated by default or early repayment or the maximum observation time of 40 months if no termination event takes place. For example mortgage A is repaid 18 months after it is issued and mortgage B defaults after 30 months. Mortgage C and F have not yet experienced a default or early repayment event at the end of the sample period of 40 months, i.e. they are right-censored observations.

Mortgage	Age borrower	default	early repayment	time
A	42	1	0	18
B	31	0	1	30
C	28	0	0	40
D	35	1	0	36
E	53	0	1	22
F	25	0	0	40

Table 6.1: Example data set

If we assume that the age of a borrower contains no information relevant for either the probability of default or the probability of early repayment, we can derive the overall probability of survival of a mortgage by applying the Kaplan-Meier estimator of equation (5.8). All mortgages were outstanding at time $t = 0$ and remain so until mortgage A defaults after 18 months. So, the estimated survival probability $\hat{S}(t) = 1$ for $t < 18$. We consider the estimate of the survival probability at exactly 18 months, the value of this

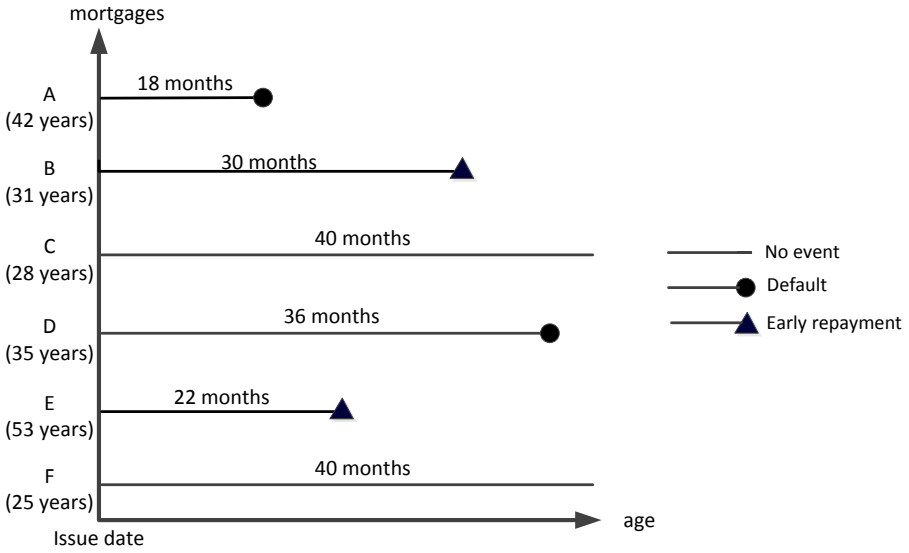


Figure 6.1: Example data set

estimate is

$$\hat{S}(18) = 1.0 \cdot [1 - 1/6] = 5/6.$$

The probability of termination in the interval $(18, 22)$ is zero and thus $\hat{S}(t) = 5/6$ for $t \in [18, 22)$. Next, we derive the estimated survival probability for $t = 22$,

$$\hat{S}(22) = 5/6 \cdot [1 - 1/5] = 2/3.$$

By continuing in the same way we find the estimated survival probability of a mortgage during the first 40 months after issuance as displayed in figure 6.2.

The example becomes more interesting when we also use the information of the age of a borrower and account for the fact that a mortgage can be terminated by two competing risks. By formula (6.4) we can write the logarithm of the likelihood function for default as

$$pl_{def}(\boldsymbol{\beta}) = \sum_{i:\delta_i=1} \left(\boldsymbol{\beta}^T \mathbf{X}_{(i)} - \ln \left[\sum_{j \in \mathcal{R}(\tau_{(i)})} \exp(\boldsymbol{\beta}^T \mathbf{X}_j) \right] \right)$$

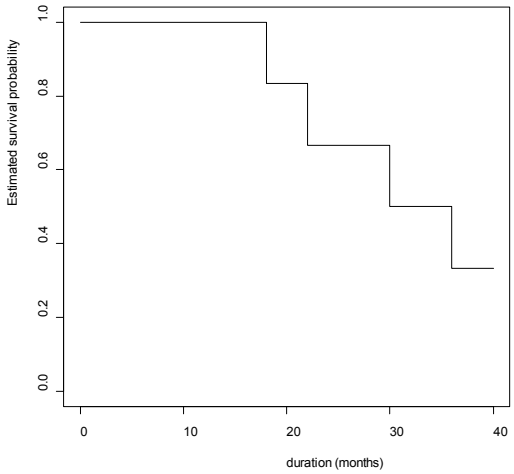


Figure 6.2: Graph of the Kaplan-Meier estimate of the survival function for the example data set

$$= \beta_{def} \cdot X_A - \ln \left[\sum_{m \in \{A,B,C,D,E,F\}} \exp(\beta_{def} X_m) \right]$$
$$+ \beta_{def} X_D - \ln \left[\sum_{m \in \{C,D,F\}} \exp(\beta_{def} X_m) \right] ,$$

where X_m is the age of the borrower of mortgage m. By maximizing this equation over β_{def}, which in this case is a vector of only one element, we find that $\beta_{def} = 0.1015$.

After obtaining β_{def} we can calculate the corresponding baseline by formula (6.8). The only default events in the data happen at time points $t = 18$ and $t = 36$, so it holds that $\hat{\lambda}_{0,def}(t) = 0$ for $t \notin \{18, 36\}$. We find that for $t = 18$

$$\hat{\lambda}_{0,def}(18) = \frac{1}{\sum_{m \in \{A,B,C,D,E,F\}} \exp(\beta_{def} X_m)} = 0.0026558,$$

and for $t = 36$

$$\hat{\lambda}_{0,def}(36) = \frac{1}{\sum_{m \in \{C,D,F\}} \exp(\beta_{def} X_m)} = 0.0154381.$$

By following the same steps for early repayment we find that $\beta_{ER} = 0.1894$

and

$$\hat{\lambda}_{0,ER} = \begin{cases} 0 & \text{for } t \notin \{22, 30\} \\ 0.000041 & \text{for } t = 22 \\ 0.000702 & \text{for } t = 30. \end{cases}$$

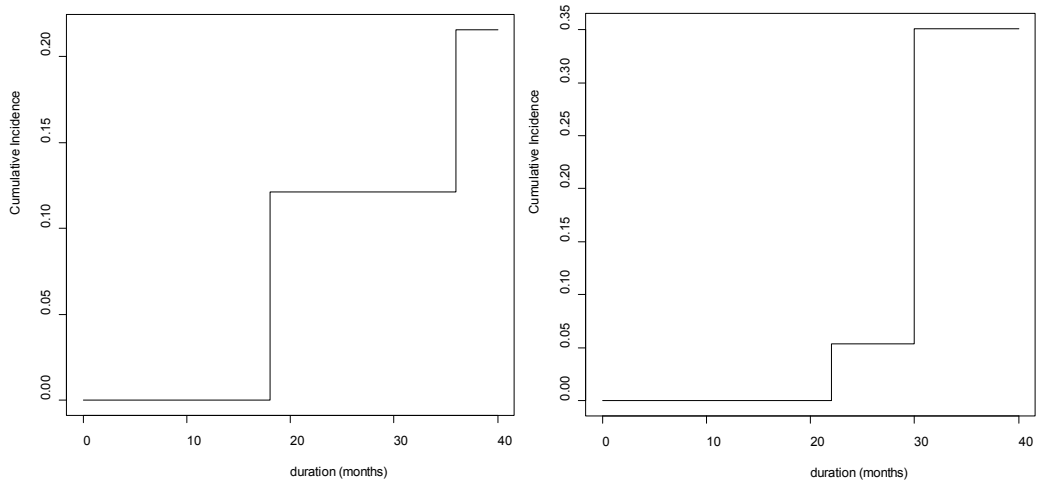
We have now fully characterized the model. We assume a new mortgage of which the borrower is 39 years old when he applies for the mortgage. For this mortgage we can derive the Cumulative Incidence Function (CIF) as defined in definition 5.6 to estimate the probability of a default or early repayment event. The CIF_{def} gives the probability of the occurrence of a default event for this new mortgage. We write

$$P\left(\tau \leq t, \epsilon = \text{default}\right) = \int_0^t \lambda_{def}(s) \exp\left(-\int_0^s \left(\lambda_{def}(u) + \lambda_{ER}(u)\, du\right)\right) ds. \tag{6.10}$$

However, since $\lambda_{0,def}(t) = 0$ for $t \notin \{18, 36\}$ it holds that $\lambda_{def}(t) = 0$ for $t \notin \{18, 36\}$. By the same reasoning we find that $\lambda_{ER}(t) = 0$ for $t \notin \{22, 30\}$, and thus we can write

$$\begin{aligned}
P(\tau < 18, \epsilon = \text{default}) &= 0 \\
P(\tau \leq 18, \epsilon = \text{default}) &= \lambda_{def}(18) \cdot \exp\left(-\lambda_{def}(18)\right) \\
&= 0.1212. \\
P(\tau < 36, \epsilon = \text{default}) &= P(\tau \leq 18, \epsilon = \text{default}) \\
P(\tau \leq 36, \epsilon = \text{default}) &= P(\tau \leq 18, \epsilon = \text{default}) + \lambda_{def}(36) \\
&\quad \cdot \exp\left(-\left(\lambda_{def}(18) + \lambda_{def}(36) + \lambda_{ER}(22) + \lambda_{ER}(30)\right)\right) \\
&= 0.2157. \\
P(\tau > 36, \epsilon = \text{default}) &= 0.
\end{aligned}$$

We have used that $\lambda_j(t) = \lambda_{0,j}(t) \cdot \exp(\beta_j \cdot 39)$ for j =default,early repayment. The cumulative probability of default is graphically displayed in figure 6.3a. The same method can be applied to the probability of early repayment to find for this same mortgage with a 39 year old borrower the cumulative probability of early repayment as displayed in figure 6.3b.



(a) Default (b) Early repayment

Figure 6.3: Cumulative probability of (a) default and (b) early repayment for working example

This example has shown in a simplified setting the methods we will apply on our much larger data set to find an estimate for the probability of default and the probability of early repayment of a specific mortgage. Note that since this example contains only six mortgages, results may seem quite odd. For example, a default event can in the model derived from the example data set only occur when the mortgage is either 18 or 36 months old. This is not an issue when the number of mortgages in the data set and especially the number of events significantly increases, as for our actual data set holds.
To clarify the methods we have assumed in this example some simplifications; besides the inclusion of time-varying covariates which was already discussed in section 5.4, we still have to discuss two more complicating factors. Firstly, the formulas (6.4) and (6.8) should be adjusted to account for possible ties in the data, we discuss this in the next section. The other aspect is that we have to account for the fact that our study is a delayed entry study to ensure that the estimated model is not biased, this is discussed in section 6.5.

6.4 Ties in the data

Our data is recorded on a monthly basis, whereas the Cox proportional hazards model is a continuous time model. This is not a real problem for data from a long sample period as ours, but it does mean that we have to account for ties in the data.

Definition 6.1. *Ties occur when at some measurement time more than one subject under observation experiences the event of interest.*

Let $\tau_{(1)} < \ldots < \tau_{(m)}$, with $m \leq k$ denote the m distinct, ordered event times. Several methods have been developed to take care of tied times; an exact expression derived by Kalbfleisch and Prentice (1980) and approximations due to Breslow (1974) and Efron (1977). Note that the Breslow approximation for the partial likelihood in the presence of ties, besides being named after the same statistician, is unrelated to the Breslow estimator for the baseline. We will not present the mathematical expression for the exact partial likelihood function here. The basis for its construction is the assumption that the d ties are due to a lack of precision in measuring survival times. The exact partial likelihood is obtained by modifying the denominator of (6.3) to include each of the d factorial arrangements of their values at each risk set. The Breslow approximation follows unmodified the approach in (6.3) even when ties are present and thus maximizes the partial likelihood function:

$$PL_B(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_{(i)+})}{\left[\sum_{j \in \mathcal{R}(\tau_{(i)})} \exp(\boldsymbol{\beta}^T \mathbf{X}_j)\right]^{d_i}}, \tag{6.11}$$

where d_i denotes the number of subjects with survival time $\tau_{(i)}$ and $\mathbf{X}_{(i)+}$ equals the sum of the covariate values over the d_i subjects. Mathematically, $\mathbf{X}_{(i)+} = \sum_{j \in D(\tau_{(i)})} \mathbf{X}_j$, where $D(\tau_{(i)})$ represents the set of indices of subjects with survival time $\tau_{(i)}$. The Efron approximation is a bit more complicated and yields a slightly better approximation to the vector $\boldsymbol{\beta}$ of the exact method in most settings, see for a simulations study comparing the Breslow and Efron approximation Hertz-Picciotto and Rockhill (1997). The

Efron approximation is an approximation for the d_i factorial possible orderings of the occurrence of events, where each ordering has equal probability:

$$PL_E(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^T \mathbf{X}_{(i)+}\right)}{\prod_{k=1}^{d_i}\left[\sum_{j\in R(\tau_{(i)})}\exp\left(\boldsymbol{\beta}^T\mathbf{X}_j\right) - \frac{k-1}{d_i}\sum_{j\in D(\tau_{(i)})}\exp\left(\boldsymbol{\beta}^T\mathbf{X}_j\right)\right]}. \tag{6.12}$$

Since the exact method is computationally too extensive for our large dataset, we will use an approximation for the tied times. The Efron method is more accurate and computationally as efficient as the Breslow method, which makes it the method of our choice.

Also for the baseline we have to account for ties in the data, see for an elaborate research on this topic Weng (2007). Since we will apply the Efron method to construct the partial likelihood, we will also apply this method to deal with the ties in estimating the baseline hazard function. For this purpose we rewrite the baseline estimator as

$$\widehat{\lambda}_{0,E}(\tau_{(i)}) = \prod_{1\le k\le d_i} \frac{1}{\prod_{k=1}^{d_i}\left[\sum_{j\in R(\tau_{(i)})}\exp(\hat{\boldsymbol{\beta}}^T\mathbf{X}_j) - \frac{k-1}{d_i}\sum_{j\in D(\tau_{(i)})}\exp(\hat{\boldsymbol{\beta}}^T\mathbf{X}_j)\right]}. \tag{6.13}$$

6.5 Delayed entry study

The standard approach in survival analysis is to define the survival time as the elapsed time from the beginning of the sample period until failure occurs. In our study we define the survival time as the elapsed time from issue date of the mortgage until default or early repayment, thus taking the age of the mortgage as the time-scale instead of the observation time of the mortgage. The main advantage of this approach is that it has a more intuitive and meaningful interpretation. Also, this approach directly takes into account the age effect on the default (or early repayment) rate.

Although the earliest mortgage in the database is issued in 1963, monitoring starts only in July 2004. Since our data consists of all mortgages outstanding in the period from July 2004 till December 2010, defining age as the time-

scale makes our study a classical delayed entry study where observations are left-truncated, see section 5.2. Left-truncation is a situation characterized by the fact that the sample does not include those subjects that have not survived long enough to be observed; for us these are mortgages issued and terminated before July 2004. Consequently, the sample that is observed is an incomplete sample and this should be taken into account in the statistical analysis. In fact, not including mortgages that have been previously terminated results in an underestimation of the failure risk, since mortgages at the highest risk are not observed.

Our sample period starts at a predetermined point in time and it is therefore reasonable to assume that the delayed entry process is independent of the survival distribution. Li (2010) describes methods to accommodate for dependently left-truncated data in survival analysis. The key for dealing in a correct manner with the delayed entries is now in the definition of survival time and the risk set at each time t in the partial likelihood formula (6.3) and the baseline estimation (6.8). Let Y be a variable measuring the exposure time, or time from entry into the study until termination or censoring, i.e. censoring occurs when the mortgage is not terminated by default or early repayment before the end of 2010 or the legal maturity of the mortgage falls in the observation period. And let W denote the delayed entry, that is the elapsed time from issue date of the mortgage until monitoring starts in July 2004. We can now write τ, the complete survival time, as the sum of these two variables, i.e. $\tau = W + Y$. Based on these quantities we can define the risk set at each time point, such that it is an unbiased estimator for the parameter $\boldsymbol{\beta}$ and the baseline function. Figure 6.4 illustrates the concept by displaying the above variables for four different mortgages, assuming that we are for the time being only interested in an event of default. Thus early repaid mortgages are treated as ordinary right-censored observations. Note that the time axis for this figure is different from that in figure 5.1, where we displayed the kind of censored observations that we have in our data. The time scale in figure 6.4 is based on the definition of the survival time, where

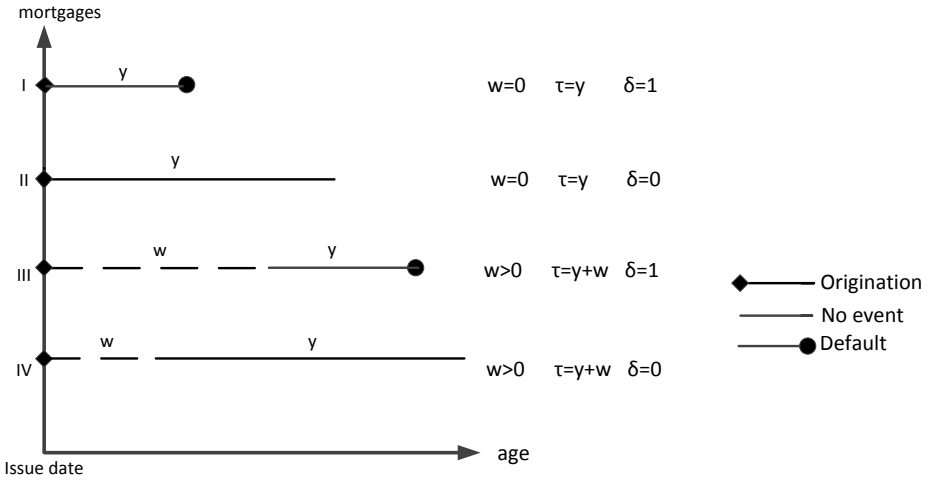


Figure 6.4: definition of survival time. y=time from entering the study until censoring or default, w=time from issue until first observation

$t = 0$ is the time of issuance of a mortgage. Mortgage I and II are issued after July 2004 and therefore enter the study at issue date. Mortgage I defaults before the end of the sample period, whereas mortgage II is still outstanding by then. Mortgage III is issued before July 2004 and therefore enters the study delayed as does mortgage IV. An important feature of the proposed method is that mortgages are not considered to be at risk prior to the age at which they enter the study. Therefore, a subject should be counted in the risk set at time t if this subject is associated with an entrance time smaller than t and a failure time greater than t. Formally we can write

$$\mathcal{R}(\tau_{(i)}) = \{j = 1, \ldots, n | W_j \leq \tau_{(i)} \leq \tau_j\}. \tag{6.14}$$

Specifically, mortgage III does not contribute to the risk set at the time that mortgage I defaults. A detailed example of this method to deal with delayed entry as well as right-censoring in survival analysis is in Lamarca et al. (1998).

Note that a related issue is the definition of a default event. As discussed in section 4.1 we consider a mortgage to have defaulted when it has been in ar-

rears for more than three months. Such a mortgage could start paying again and thereby becomes once more a performing mortgage; in this situation the mortgage will after starting to pay again be, as before, part of the risk set. This is not the case for prepaid mortgages which will obviously be removed from the risk set the moment there is no outstanding balance on the loan any more.

Chapter 7

Characteristics of data set and model development

This chapter describes the available data in the first section. The second section gives an outline of the model development steps.

7.1 Characteristics of data set

NIBC has maintained a database of all Dutch residential mortgages on the balance of the bank between 01/07/2004 and 31/12/2010. The data is recorded on a monthly basis and contains 3,350,022 observations of 70,518 mortgages. The mortgages are issued between February 1963 and December 2010 and 1,760 defaults and 33,408 early repayments have been recorded. In table 7.1 a summary is given of the variables in the study. Some of these data points are recorded at time of issuance of the mortgage, while others are calculated based on given data or extracted from the market. Also note that some variables are static, while others are time-varying. Since the distinction is not always completely intuitive, the time-varying variables are indicated by a T in the table. For example, even though the income of a borrower might change over the maturity of the mortgage, it is only recorded at issue date, making it a static variable. Some of the variables need some further

Variable	Description	Codes/Values	Time-varying
ID	Identification code		
IsDate	Issue date	dd/mm/yyyy	
MatDate	maturity date	dd/mm/yyyy	
ObDate	month of observation	mm/yyyy	T
OutB	outstanding balance	Euro's	T
VProp	Value of the property	Euro's	
Inc	Yearly income of borrower(s), there may be 1-3 reported incomes	Euro's	
NumApp	Number applicants [1]		
Int	interest on the mortgage loan	percentage	T
AdV	Advisor verified	0=No,1=Yes	
BKR	negative credit history	0=No,1=Yes	
Age	age of (oldest and youngest) borrower	years	T
Reg	region of the underlying property	Zeeland,Utrecht,etc	
IRDate	interest reset date	0=No,1=Yes	T
LTFV	ratio of outstanding loan to foreclosure value [2]	percentage	T
LTiFV	ratio of outstanding loan to indexed foreclosure value [3]	percentage	T
IPTI	ratio of monthly interest payments to monthly total income	Euro's/Euro's	T
LTI	total loan amount divided by total yearly income	Euro's/Euro's	T
SMI	largest income divided by total income [4] associated with the mortgage	percentage	
3ME	3-months Euribor rate	percentage	T
5YS	5 year versus 3 months swap rate	percentage	T
10YS	10 year versus 3 months swap rate	percentage	T
RetSpr	Retail spread	percentage	T
RefInc	Refinancing incentive		T

[1] although the majority of mortgages is held by 1 or 2 applicants, values up to 10 are registered.

[2] in Dutch 'executiewaarde'.

[3] this index is the house price index reported monthly by the 'Kadaster'

[4] A maximum of 3 incomes may be registered for each loan. This variable is a measure of resiliency to unemployment.

Table 7.1: Description of variables in mortgage data

explanation:

- Advisor verified: an income is advisor verified if the borrower does not disclose his income to the lender, but to an intermediary.

- BKR: a Dutch institution registering credit history of inhabitants. A '1' indicates the person has been in default or in arrears of any financial obligation to a firm reporting to BKR in the past, a '0' means there is no such information.

- Interest reset date: at such a date the borrower can refinance his mortgage in the market without paying the prepayment penalty, which can be a significant amount. This variable only plays a role in modelling early repayment, not in modelling default.

- 3 months Euribor: a reference rate quoted daily at which banks offer to lend unsecured funds to other banks in the inter-bank market for a period of three months. The interest payment of variable rate mortgages is linked to the 3-months Euribor plus some spread to cover risk and expenses.

- The 5 year versus 3 months swap: a derivative in which one party receives every month 3-months Euribor, fixing every 3 months, and pays a fixed interest rate for the maturity of 5 years. This fixed rate is quoted daily at the market such that the arbitrage free value of the swap at issue date is zero. This quote plus some spread is the basis for the adjustable rate mortgages. Equivalent for the 10 year versus 3 months swap.

- Retail spread: measures the average spread over the swap curve that is charged in the mortgage market. This spread captures operational risk, credit risk and funding costs. Although this variable should not be interpreted as the state of the macro economy, many factors regarding mortgage credit risk are captured in this variable.

- Refinancing incentive: is a measure for the incentive to refinance a mortgage in the market. When the borrower pays a high interest rate while market quotes are low, the incentive to refinance will be relatively high. It is defined as the interest paid by the borrower divided by an adjusted market interest rate. This market interest rate is mortgage specific and reflects the interest the borrower would have to pay if he decides to refinance his debt and it is calculated by the 5 year versus 3 months swap rate from the market plus the retail spread plus or minus some basispoints reflecting the risk category the specific mortgage belongs to. If we call this last part x, we can write

$$\text{Refinancing incentive} = \frac{\text{Interest on the mortgage}}{\text{5Y vs 3M swap rate} + \text{retail spread} + x}$$

7.2 Model development

We will carry out the statistical analysis of the data in the software program R (version 2.10.0), which has some very useful packages available. We use the package "Survival" to estimate the influence of the covariates and the form of the baseline. Dedicated code was written for data preparation and estimating the cumulative incidence function by the approach of Kalbfleisch and Prentice.
In this section we discuss the development of the model. For this purpose we will first describe some statistics to analyse the performance and significance of a model. The next subsection describes the actual steps in the process of choosing the covariates in the model. And the final subsection gives the details of one of these steps, namely methods to analyse the scale of a continuous covariate.

7.2.1 Assessment of model significance

Typically, the first step following the fit of a regression model is the assessment of the significance of the model and the model covariates. The relevant

models for assessing the significance of a covariate are the partial likelihood ratio test, the Wald test and the score test, which can all be obtained directly from R. The null hypothesis for all statistics is that the coefficient is equal to zero. The partial likelihood ratio test is calculated as

$$G = 2\left(L_p(\widehat{\beta}) - L_p(0)\right), \tag{7.1}$$

where $L_p(\widehat{\beta})$ is the log partial likelihood of the model containing the covariate and $L_p(0)$ is the log partial likelihood for the model not containing the covariate. Under the null hypothesis this statistic will follow a chi-square distribution with 1 degree of freedom. The Wald statistic is defined as

$$Z = \frac{\widehat{\beta}}{\widehat{SE}(\widehat{\beta})}, \tag{7.2}$$

where $\widehat{SE}$ is the estimated standard error. This Wald statistic follows under the null hypothesis a standard normal distribution.
The score test is a third test frequently used. The equation for the score test is

$$S = \left.\frac{\partial L_p/\partial\beta}{\sqrt{\mathbf{I}(\beta)}}\right|_{\beta=0}, \tag{7.3}$$

where

$$\mathbf{I}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial\beta^2}. \tag{7.4}$$

This test also follows the standard normal distribution under the null hypothesis. In practice, the same conclusion is usually drawn from the three tests about the significance of the coefficient. In situations where there is disagreement, we will choose to use the Wald statistic.

7.2.2 Purposeful selection of covariates

For the model building process we will follow the approach outlined by Hosmer et al. (2008, p. 133-136), which we can summarize as follows:

1. We begin by fitting a multivariable model containing all variables which are significant in the univariable analysis at the 25% level by the Wald test.

2. We remove all covariates that are not significant as identified by the Wald test of the individual covariate one by one. At this stage we take a significance of 5% to ensure we do not delete too many variables at once.

3. Based on the fit of the reduced model, we assess whether the removal of the covariate has caused a change of more than 20% of the coefficients of the remaining covariates. If this is the case we add the covariate back into the model.

4. When no more covariates can be removed, we add to the model one by one the covariates initially excluded from the multivariable model to confirm that they are not statistically significant in the presence of the remaining covariates.

5. We check for each continuous covariate in the resulting model the linearity of the log hazard (the proportional hazards assumption) and when necessary transform the continuous covariate. See for details of this step the next subsection.

6. We determine whether there are interaction terms needed in the model.

7. The final step is the evaluation of the model for overall goodness-of-fit and checking the model assumptions.

7.2.3 Methods to examine scale of continuous covariates

An important modelling step is to determine whether the data supports the assumption of linearity in the log hazard for all continuous covariates. A common practice in medical studies is to convert continuous covariates into

binary dummy variables. Although such a model may not be optimal for continuous covariates, the decision to use such a model is often made on the grounds that it is easier to interpret the results in case of binary covariates (Klein and Wu, 2004). As Royston and Sauerbrei (2008) mention, categorization introduces the problem of defining cut point(s), overparametrisation and loss of efficiency. They state that in any case, a cut point model is an unrealistic way to describe a smooth relationship between a predictor and an outcome variable. An alternative approach is to keep the variable continuous and allow for some form of nonlinearity. We discuss in this section two methods that can be performed to analyse the assumption of linearity in the log hazard and suggest possible transformations.

The simplest method is to replace the covariate with design variables from its quartiles, where the first quartile is used as a reference. If plotting the estimated coefficients for the design variables against the midpoints of the intervals gives an approximately straight line, the scale can be assumed linear in the log hazard. If the line connecting the points substantially departs from a linear trend it might suggest a transformation of the covariate.
A second more advanced method is the fractional polynomials method. This method can be used with a multivariable regression model by applying the method to the continuous variables one after another (or even iteratively). However, for simplicity we describe the method here for a model with a single continuous covariate. We generalize the hazard function to

$$\lambda(t|X) = \lambda_0(t) \cdot \exp\left(\sum_{j=1}^{J} F_j(x)\beta_j\right) , \tag{7.5}$$

where $F_j(x)$ is a particular type of power function. Although, we could allow the covariate to enter the model with any number of functions, we will restrict the transformation to a maximum of two powers. The value of the first function is $F_1(x) = x^{p_1}$, and that of the second is defined as

$$F_2(x) = \begin{cases} x^{p_2} & \text{if } p_2 \neq p_1 \\ F_1(x) \cdot \ln(x) & \text{if } p_2 = p_1 \end{cases}$$

In theory the power, p_j, could be any number, but for practical purposes Royston and Altman (1994) propose to restrict the power to be among those in the set

$$\Omega = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\},$$

Where $p_j = 0$ denotes the natural log of the variable.

The steps and methods discussed in this section are carried out in the appendices B.1 and B.2 to obtain a default model and an early repayment model respectively. The results of these steps are presented in the next chapter.

Chapter 8

Results

In this chapter we describe the results of the estimated default and early repayment models as well as the simulation process. The first two sections describe respectively the default and early repayment model, discuss the interpretation and give some theoretical justification of the covariates in the model. The third section describes our practical approach in modelling the LGD and the final section discusses the assumptions we made in the simulation process and some simulation results for a specific RMBS.

8.1 Default model

By following the steps outlined in section 7.2 we obtained a model to describe the probability of default for a specific mortgage. The covariates influencing the probability of default and the corresponding maximum likelihood estimates for the β coefficients are displayed in table 8.1. The details of the model fitting process can be found in appendix B.1.
From table 8.1 we see that all parameters have a significant influence on the probability of default and we can write the model as follows:

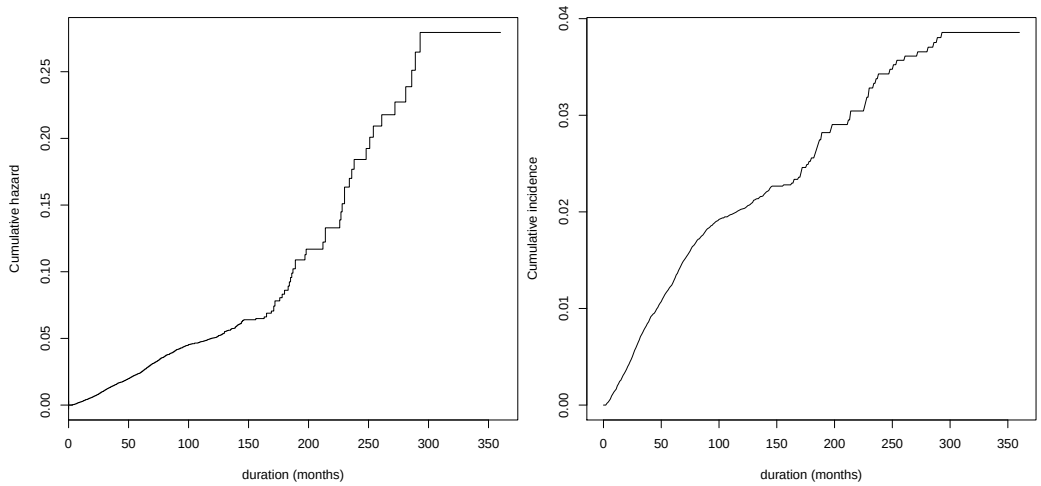
$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp\Big(0.412748 \cdot \sqrt{\text{LTFV}} + 1.872541 \cdot \text{BKR} + 0.013006 \cdot \text{SMI} - 0.185020 \cdot \text{NumbApp}_{\{0,1\}} + 1.481644 \cdot \text{AdV}\Big). \quad (8.1)$$

	Coefficient	exp(coef)	se(coef)	z	P-value
$\sqrt{\text{Loan-to-Foreclosure-Value}}$	0.412748	1.510964	0.028906	14.279	<2e-16
BKR	1.872541	6.504805	0.138789	13.492	<2e-16
Share main income	0.013006	1.013090	0.002266	5.739	9.51e-09
Number applicants$_{\{0,1\}}$	-0.185020	0.831087	0.080746	-2.291	0.0219
Advisor verified	1.481644	4.400173	0.094783	15.632	< 2e-16

Table 8.1: Default model

We will now discuss the interpretation of the model. Starting with the BKR covariate, we can interpret the β as follows: ceteris paribus, borrowers with a negative credit history (BKR) default at a 5.5 times higher rate than borrowers without a negative credit history. In the same way, we can interpret the results for Advisor verified and Number applicants. Where for Number applicants we have to remark that we made this variable into a binary variable, with a '0' indicating that there is only one borrower and a '1' that there are two or more applicants. This means that mortgages with two or more applicants default at a rate 17% lower than mortgages on one name. Share main income (SMI) is a continuous covariate for which we can interpret the result as follows: a 1 percentage point increase in the SMI, for example from 50% to 51%, results in a 1.3% higher rate of default. And the estimated hazard ratio for a 10 percentage point increase of the SMI is $\exp(10 \cdot 0.013006) \approx 1.139$. This means for example that mortgages with a SMI of 100% default at a 14% higher rate than mortgages with a SMI of 90%. The interpretation of the value for the LTFV is a little less intuitive, since we applied a nonlinear transformation to this value. For example, borrowers with a LTFV of 110% having all other covariates equal, default at a 22% higher rate than borrowers with a LTFV of 100%, i.e. $\exp\left(\left(\sqrt{110} - \sqrt{100}\right) \cdot 0.412748\right) \approx 1.2232$.

Figure 8.1a displays the cumulative hazard for a mortgage with the following characteristics: a LTFV of 100, Share main income of 100, 1 applicant, no BKR and non-advisor verified. The straight line from a duration of 300 months on indicates that in the data there are no defaults recoreded of mort-



(a) Cumulative hazard (b) Cumulative probability

Figure 8.1: Graphs of (a) cumulative hazard and (b) cumulative incidence function of default

gages older than 300 months. Also it seems like the cumulative hazard is rapidly increasing after 150 months which would indicate an increasing probability of default after 150 months. Unfortunately, the interpretation of the cumulative hazard in the presence of a competing risk is not that straightforward. As discussed in subsection 5.5.2 we cannot conclude anything about the probability of default from the hazard rate without accounting for the probability that a mortgage is repaid. Figure 8.1b displays the Cumulative Incidence Function for the same mortgage, which is defined as

$$F_{\text{default}}(t) = P\left(\tau \leq t, \epsilon = \text{default}\right) . \tag{8.2}$$

The CIF can be interpreted as the cumulative probability of default in the presence of the competing risk of terminating a mortgage by early repayment. For the calculation of the CIF use is made of formula (5.14) and also the early repayment model is needed as input; this model is further discussed in the next section. Based on the CIF we can calculate the probability that a certain mortgage with specific age and characteristics will default in the next month. For example, if we take again the same mortgage, and assume it is issued

one year ago, the 1-month probability of default, according to the model, is

$$\begin{aligned} &P\left(\tau = 13, \epsilon = \text{default} | \tau > 12\right) \\ &= \frac{P\left(\tau \leq 13, \epsilon = \text{default}\right) - P\left(\tau \leq 12, \epsilon = \text{default}\right)}{P(\tau > 12)} \\ &= \frac{0.002000 - 0.001652}{0.968527} = 0.000363. \end{aligned}$$

Equivalently, for this same mortgage, if it is still outstanding after 10 years, its probability of default in the next month is equal to $\frac{0.020296-0.020264}{0.316971} = 0.000102$.

Some theoretical justification for the covariates in the model is as follows:

- LTFV. A higher Loan-to-foreclosure-value at origination indicates that the borrower provided little or no own equity for his property. Thereby the borrower has less incentive to continue paying the mortgage debt.

- BKR code. A borrower that has defaulted or has been in arrears on a financial obligation in the past is more likely to default on his mortgage loan.

- Share main income. Unemployment of borrowers is a key driver in the credit risk of residential mortgages. This risk may be partially mitigated if the mortgage loan is associated with more than one income.

- Number of applicants. The number of applicants is also associated with the ability to cushion unemployment.

- Advisor verified. Intuitively it is not surprising that advisor-verified loans have a higher PD than non-advisor-verified loans, since there is a higher uncertainty of income-statements for these loans and these borrowers are often entrepreneurs whose income is more volatile. The risk is partly mitigated by the lower LTFV of advisor verified-loans.

8.2 Early repayment model

The covariates influencing the probability of early repayment and the corresponding maximum likelihood estimates for the β coefficients are displayed in table 8.2. The details of the intermediary steps can be found in appendix B.2. The final model can be written as

$$\begin{aligned}\lambda(t|\mathbf{X}) = \lambda_0(t)\cdot \exp\Big(&- 0.573927 \cdot \ln(\text{LTFV}) + 0.010762 \cdot \text{LTFV} \\ &-0.008309 \cdot \text{Age} + 1.858693/\sqrt{\text{Inc}} + 0.154317 \cdot 3\text{ME} \\ &+1.009755 \cdot \text{IRDate} + 1.044252 \cdot \text{RefInc}\Big). \end{aligned} \tag{8.3}$$

	Coefficient	exp(coef)	se(coef)	z	P-value
ln(Loan-To-Foreclosure-Value)	-0.573927	0.563309	0.034304	-16.731	<2e-16
Loan-To-Foreclosure-Value	0.010762	1.010820	0.000531	20.270	<2e-16
Age youngest applicant	-0.008309	0.991726	0.000543	-15.313	<2e-16
$1/\sqrt{\text{Income}}$	1.858693	6.415348	0.069052	26.917	<2e-16
3-months Euribor	0.154317	1.166860	0.004072	37.897	<2e-16
Interest reset date	1.009755	2.744928	0.018534	54.483	<2e-16
Refinancing incentive	1.044252	2.841272	0.021116	49.453	<2e-16

Table 8.2: Early repayment model

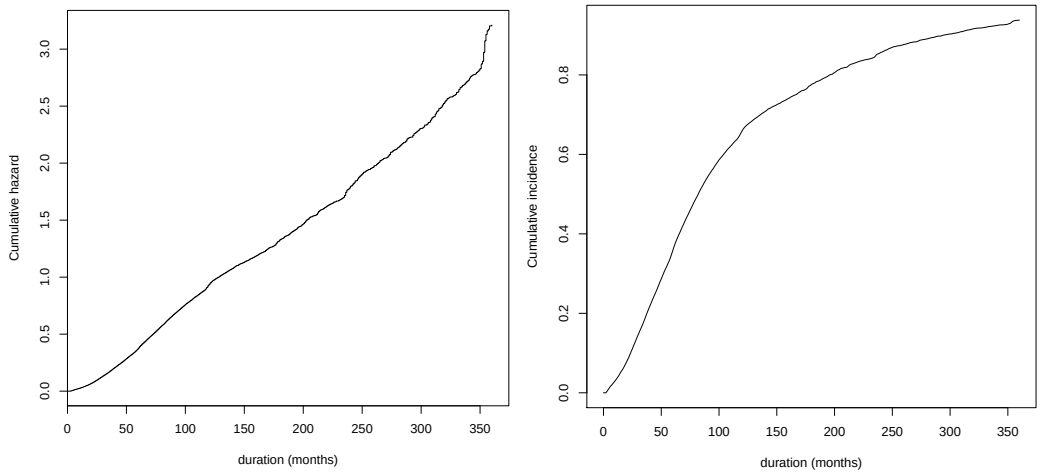
Although the interpretation of the model is similar to the interpretation of the PD model discussed in the previous section, we will briefly discuss it for the ER model too, to get some comfort with the model. An increase of the age of the youngest applicant has a decreasing effect on the rate of early repayment. The estimated hazard ratio for a 10 year increase of the age of the youngest applicant is $\exp(10 \cdot -0.008309) \approx 0.92$. This means that mortgages for which the youngest borrower is 10 years older early repay at a 8% lower rate than younger borrowers. For Euribor the effect of an increase is positive and can be interpreted in the same manner, thus a one procent point increase gives an almost 17% higher rate of early repayment. For the refinancing incentive again the interpretation is equivalent

and a one procent point increase gives a 1.84 times higher rate of early repayment. Reset date is a binairy variable for which we can see from table 8.2 that the rate of early repayment at a reset date is 1.74 times higher than at another date, all other things being equal. For both LTFV and Income the interpretation is harder, since their effect is not linear. For example, borrowers with a LTFV of 120% having all other covariates equal, early repay at an almost 12% higher rate than borrowers with a LTFV of 100%, i.e. $\exp\left(-0.573927 \cdot (\ln(120) - \ln(100)) + 0.010762 \cdot (120 - 100)\right) \approx 1.117$. For Income, which is measured in thousands of euros, there is a negative correlation between the rate of early repayment and income. The average income is about €42,000 a year. Borrowers with a yearly income of €30,000 early repay at an almost 8% higher rate than, ceteris paribus, borrowers earning €50,000 a year, i.e. $\exp\left(1.85869 \cdot \left(1/\sqrt{30} - 1/\sqrt{50}\right)\right) \approx 1.079$.

Figure 8.2a displays the cumulative hazard for a mortgage with a LTFV of 100, the age of the youngest applicant is 41, an income of 42000 and Euribor is constant at 2.747% while the refinancing incentive is constant at 1.092; these are average values. Figure 8.2b displays the cumulative incidence function of early repayment of this same mortgage for which we also know that there is no negative BKR registration, it is non-advisor verified and there is one applicant. This extra information is needed since the CIF also depends on the hazard rate of default. The CIF can be interpreted as the cumulative probability of early repayment, see section 5.5.2 and 8.1 for more details. For example, if the mortgage described above was issued 5 years ago, we find a 1-month probability of early repayment (ER) of

$$\begin{aligned} &P\left(\tau = 61, \epsilon = \text{ER} | \tau > 60\right) \\ &\quad = \frac{P\left(\tau \leq 61, \epsilon = \text{ER}\right) - P\left(\tau \leq 60, \epsilon = \text{ER}\right)}{P(\tau > 60)} \\ &\quad = \frac{0.358136 - 0.349219}{0.627635} = 0.014207. \end{aligned}$$

And when this mortgage is still outstanding after 10 years, the 1-month probability of early repayment is equal to 0.016354.



(a) Cumulative hazard (b) Cumulative probability

Figure 8.2: Graphs of (a) cumulative hazard and (b) cumulative incidence function of early repayment of a mortgage

Some theoretical justification for the covariates in the model is as follows:

- LTFV. Following the reasoning of Alink (2002), the LTFV is in the Netherlands not an indicator of wealth as it is in other countries. It can be seen as an indicator of financial awareness. Because interest is tax deductible in the Netherlands, it is advantageous for borrowers to have a high as possible mortgage loan. Therefore we assume that borrowers who are more financially aware will take out loans with higher LTFV's. When we take this together with Alink's proposition that more financial aware borrowers will repay faster, we can conclude that a higher LTFV leads to higher early repayment rates.

- Age youngest applicant. Young borrowers typically have lower incomes and when they start to get children and their income increases they will look around for another house. Also younger people tend to move more often, resulting in higher repayment rates.

- Income. Since borrowers with a high income have easier access to

refinancing opportunities, we would expected a positive correlation between the probability of early repayment and income. However, from our analysis it turned out that income is negatively correlated with the probability of early repayment. This was also the result of the univariable analysis and we are therefore comfortable that we are not overfitting the data. One possible explanation is that borrowers with higher incomes can profit more from tax savings by reducing early repayment of their mortgages.

- Euribor. This variable is the rate at which banks offer to lend unsecured funds to other banks for a 3-months period, it is an indicator for the overall market conditions.

- Interest reset date. By refinancing a mortgage at an interest reset date the borrower does not have to pay the lender a prepayment penalty. This can be a significant amount and thus it is for the borrower attractive to refinance his mortgage at an interest reset date, instead of at any other date.

- Refinancing incentive. This variable is a measure for the incentive to refinance a mortgage in the market. It is defined as the interest paid by the borrower divided by an adjusted market interest rate. The refinancing incentive is bigger than one when a borrower pays a higher interest rate on his mortgage than he would when applying for a new loan. By prepaying the existing mortgage and taking out a new mortgage loan, he could have an economic benefit.

The interested reader is referred to Alink (2002) for details on early repayments in the Netherlands.

8.3 LGD model

The LGD is the incurred loss in case a borrower defaults on his mortgage. This amount will depend on the outstanding balance of the loan and the

value of the underlying property. In the years before the credit crisis house prices were rising rapidly, making actual losses for lenders a rare event. If a borrower was not able to pay the mortgage instalments, the proceeds from the foreclosure process were almost always enough to cover the mortgage debt. We can conclude from this that the market value of a property instead of the value as registered at issue date of the mortgage should be considered in determining the LGD. An approximation for the market value can be found by using the house price index as published by 'het kadaster'. However, since house prices exhibit a lot of autocorrelation, a good prediction of future house prices is highly complicated. Also the liquidity in the market might influence the proceeds from the foreclosure process. From the credit crisis it became evident that even though a house is in good shape it might be unsaleable due to an illiquid market. Although the incurred loss in case of a residential mortgage has a significant impact on the cash flows to the RMBS notes, an in-depth research on the LGD is outside the scope of this thesis.

Instead, we will approach the LGD issue from a more practical point of view and exploit the knowledge on this topic present within the bank. From practice it turns out that about 60% of the defaulted mortgages starts paying again and these mortgages will eventually pay down the missed interest payments. We assume therefore that a defaulted mortgage, irrespective of it's individual characteristics, has a probability of 0.6 of becoming a performing mortgage again. For those mortgages for which the bank actually starts a foreclosure process, we split up the loss in a fixed part and a part that depends on the outstanding balance and the value of the property. This fixed part is estimated to be about € 5000, which covers administration and processing cost. For the variable part we consider the difference of the outstanding balance and the value of the property in a foreclosure process, where we will also take an extra safety factor of 10% into account. This safety factor is the factor with which we assume that we have overestimated the foreclosure proceeds based on the LTFV. For a mortgage which is taken

out under NHG conditions (in Dutch 'Nationale Hypotheek Garantie') the variable amount is taken to be only a fourth of the loss for a mortgage not taken out under NHG conditions. The reason is that the mortgage guarantee fund that operates NHG provides safety for the borrower to the lender. In formula we can write the loss as

$$\begin{aligned}\text{loss} = X + OB \cdot \max\left(0, 1 - \frac{1}{LTFV \cdot (1+SF)}\right) \cdot \mathbb{I}\{\text{not NHG}\} \\ + OB \cdot \max\left(0, 1 - \frac{1}{LTFV \cdot (1+SF)}\right) \cdot \mathbb{I}\{\text{NHG}\} \cdot 0.25\,, \quad (8.4)\end{aligned}$$

where X is the fixed amount, SF is the safety factor and OB is the outstanding balance.

8.4 Simulation

In this section we will discuss the tool developed to price RMBS notes and the results of the simulation described in chapter 3. The first subsection discusses the assumptions on which the simulation is based and the second subsection shows and discusses some results obtained for a specific transaction. We used Delphi® (Embarcadero; San Francisco, CA, USA) integrated development environment, running within the Windows operating system, to develop a valuation tool for RMBS notes. This resulted in a user-friendly application giving the user the opportunity to select the input file containing the data on the underlying mortgages, choose the desired output and define the specifics of the transaction. The tool offers several options related to the structure of the tranches, the interest swap, payment frequency, the principal and interest waterfall and technicalities with respect to the valuation method, see for more details on the developed tool appendix C.

8.4.1 Underlying assumptions

As already outlined in chapter 2 an RMBS is a highly complex investment product, coming in many forms and shapes. A prospectus for this product

will average about 150 pages of legal language, making it difficult to really understand the structure of the transaction. For this reason we choose one specific transaction of which a lot of in-house knowledge is available within the bank; called DMBS XV (Dutch Mortgage Backed Security). We will discuss now the assumptions made in the development of the RMBS pricing tool:

- Prepayment penalties are not taken into account, i.e. they are zero.
- Partial redemption of the mortgage is non existing; the only prepayments are full redemptions.
- There is no timing difference between payments made by the mortgage borrowers and the payments to the noteholders, thus we do not explicitly model a liquidity facility.
- There is no replenishment period and substitution is not explicitly modelled.
- A foreclosure process always takes the same number of months; this is a choice made by the user. We will assume for DMBS XV a period of 18 months.
- There are always sufficient funds to pay fees and other expenses.
- In case a defaulted mortgage recovers, we assume that the future payments eventually equate to those if no default had occurred. For this reason, we will also assume that the mortgage in that case recovers immediate.
- A transaction is always called at the first optional redemption date (FORD), so no step-up margin is taken into account.

We will use an interest curve based on market quotes at the issue date increased by a discount spread to define the discount curve. This discount

spread is an indication for the funding cost of the bank; it consists of a return for the risk a borrow bears when lending money to NIBC plus a liquidity premium. The discount curve is a tool to calculate the fair value of an RMBS note at issue date.

As the previous section showed, also the probability of early repayment depends on the interest rate in the market, through two variables: 3-months Euribor and repayment incentive: which value depends on the 5-year swap rate and the retail spread. We could simulate for these variables a possible path based on an interest model, however since we already obtained an interest curve which is used for valuation purposes in the model, it is more consistent to use this same interest curve in determining the values for these variables. For this purpose we can derive the 3-months and 5-year forward rate from the interest curve on any payment date after the issue date. While interest rates are quoted in the market, there is no indication for the market's expectation of the retail spread. Remember that the retail spread can be thought of as a spread over the risk free rate charged in the market to cover expenses and risks associated with a mortgage loan. As figure 8.3 indicates, the retail spread has been highly volatile during the last years.

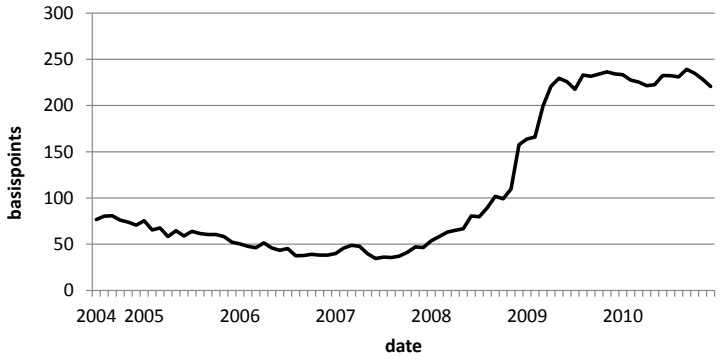


Figure 8.3: Graph of retail spread in the market in the period of 2004 till 2010

Since (1) there is no literature on the distribution of the retail spread, (2) we have only a few data points and (3) the data mainly covers the period of the financial crisis, which is not very representative for any other period in time,

we will not use a sophisticated model to derive possible paths for the retail spread. Instead we will simply model a possible path of the retail spread in the following way:

$$RS_t = (1 + Z) \cdot RS_{t-1}\,, \tag{8.5}$$

where RS_t is the retail spread at time t and Z is a normal distributed random variable, i.e. $Z \sim N(\mu, \sigma^2)$. From historical data we could estimate the mean and variance of the relative change in the retail spread and we found that $\mu = 0.01766$ and $\sigma = 0.091114$. In appendix D ten different realisations of the retail spread based on this model are displayed.

8.4.2 Results DMBS XV

DMBS XV is a 750 million Euro securitisation transaction, originated in the Netherlands and issued by NIBC in March 2010. Table 8.3 gives a summary of the notes issued in this transaction.

Note class	Amount (Size (€) [1]	Credit enhancement	Coupon [2]
Class A1	182,100,000	5.00 %	1M + 110 bps [3]
Class A2	530,600,000	5.00 %	1M + 150 bps
Class B	11,200,000	3.50 %	1M + 200 bps
Class C	10,450,000	2.10 %	1M + 300 bps
Class D	10,400,000	0.70 %	1M + 400 bps
Class E	1,500,000	0.50 %	1M + 450 bps
Class F (reserve account)	3,750,000	0.00 %	1M + 500 bps

[1] including the reserve account (F notes)

[2] 1M is 1 month Euribor rate

[3] bps stands for basis points, and it equals a one-hundredth of a percentage point

Table 8.3: DMBS XV notes

The collateral pool for this mortgage consists of 4,180 mortgages and has a size of €746,250,000. The proceeds of the notes A1 till E are used to fund the

mortgages, while the proceeds of the F notes fund the reserve account upfront. Furthermore we can summarize the characteristics of this transaction as follows:

- All notes are issued in denominations of €50,000.
- Issue date is March 25, 2010.
- The first optional redemption date is April 2, 2015.
- Payments are made monthly and the first payment date is May 3, 2010.
- Losses are recorded on the PDL of the corresponding tranche.
- The interest swap guarantees an excess spread of 50 bps a year. It is applied sequentially to absorb missed interest on the mortgages, cover losses through the PDL's, replenish the reserve fund to it's target level and pay out the remaining amount to the issuer.

Based on the funding cost of NIBC we will apply a discount spread of 150 basispoints over the risk free interest curve. We obtain for notes in tranche A2 a value exactly at par, due to the fact that these notes pay coupons of one month Euribor plus 150 basispoints, exactly the same as used in discounting. In other words, as long as this tranche is not suffering any losses, it will always value at par. There are only very minor differences in the outcomes of the 100,000 simulation runs we have performed, especially tranche B till E generate exactly the same cash flows in all runs, see table 8.4 for a summary of the obtained results.
The reason that the outcomes are identical is that there is in every realisation enough excess spread and money on the reserve account to absorb all losses, while repayment is never high enough to redeem even the smallest part of the tranches B till E before the FORD. Tranche A1 is affected by the number of borrowers early repaying their mortgage, see figure 8.4 for the cumulative discounted cash flows to tranche A1 where the lower bound (upper bound) refers to the realisation with the lowest (highest) total value

Tranche	Minimal value	Average value	Maximal value
A1	€49593.42	€49646.39	€49702.49
A2	€50000.00	€50000.00	€50000.00
B	€51106.53	€51106.53	€51106.53
C	€53305.24	€53305.24	€53305.24
D	€55485.11	€55485.11	€55485.11
E	€56568.10	€56568.10	€56568.10

Table 8.4: Result of simulation for DMBS XV

for a note. Figure 8.5 displays the corresponding (not discounted) cash flows at each monthly payment date.
Figure 8.6 displays for each payment date the minimal, maximal and median cash flow from the 100,000 simulation runs. Note that these do not correspond to an actual realisation, since the minimum, maximum and median are taken per month. Given that the tranches B till E give a very similar pattern we only display the cumulative discounted and (not discounted) cash flows per payment date for tranche E in respectively figure 8.7 and figure 8.8.

The purpose of this research was to be able to quantify the probability distribution of the value of the notes. However, since tranche A1 will always result in a loss due to the discount spread that is higher than the margin an investor in a A1 note receives, no investor will in these circumstances invest in a tranche A1 note and there is little use in calculating a probability distribution for the loss. For tranche A2 till E there is only one possible realisation and a probability distribution is also of little meaning. If we now assume that NIBC is able to fund itself against Euribor plus only 90 basispoints, we can calculate a probability distribution for the value of notes in tranche A1 and A2. We again performed 100,000 simulation runs and the resulting profit as a percentage of the initial outlay for notes in tranche A1 and A2 are displayed in respectively figure 8.9 and figure 8.10.

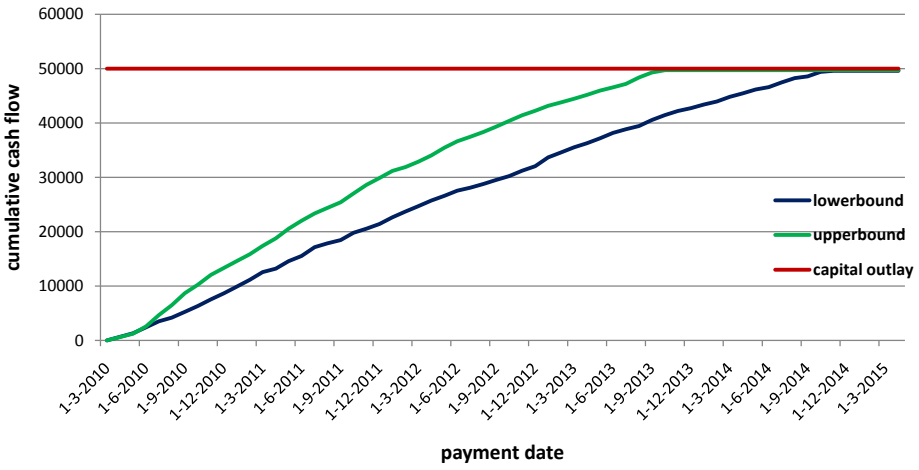


Figure 8.4: Graph of cumulative discounted cash flows to a note in tranche A1, where the upper bound refers to the realisation with the highest value and the lower bound to the realisation with the lowest value

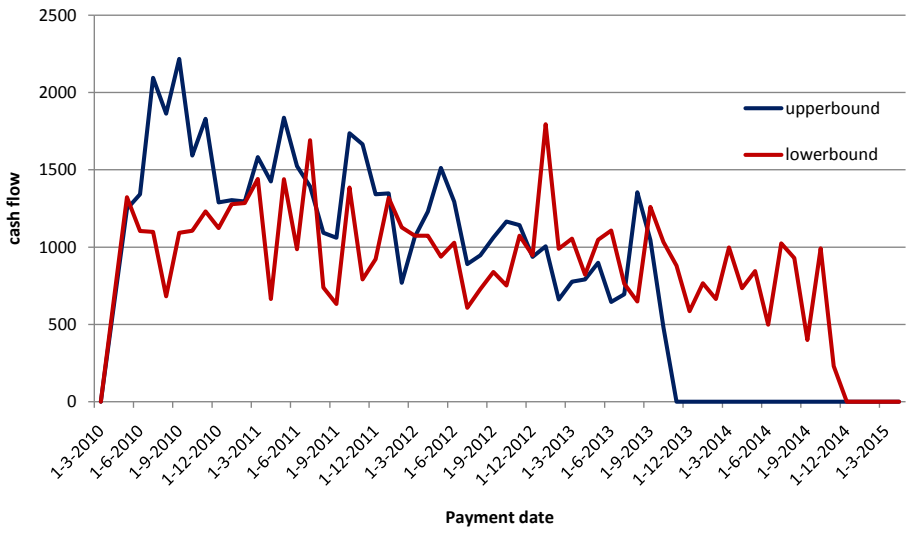


Figure 8.5: Graph of cash flows to a note in tranche A1, where the upper bound refers to the realisation with the highest value and the lower bound to the realisation with the lowest value

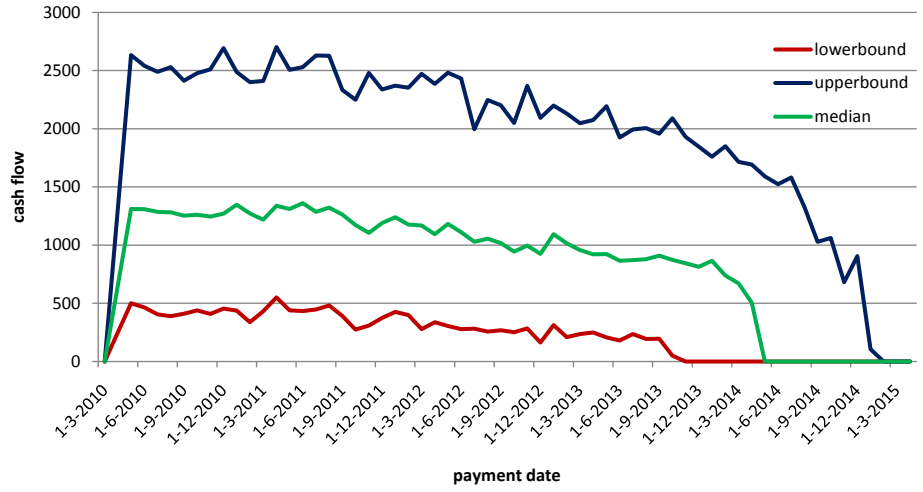


Figure 8.6: Graph of realisations of cash flows to a note in tranche A1, where each line (upper bound/lower bound/median) refers to the highest/lowest/median cash flow for that specific month

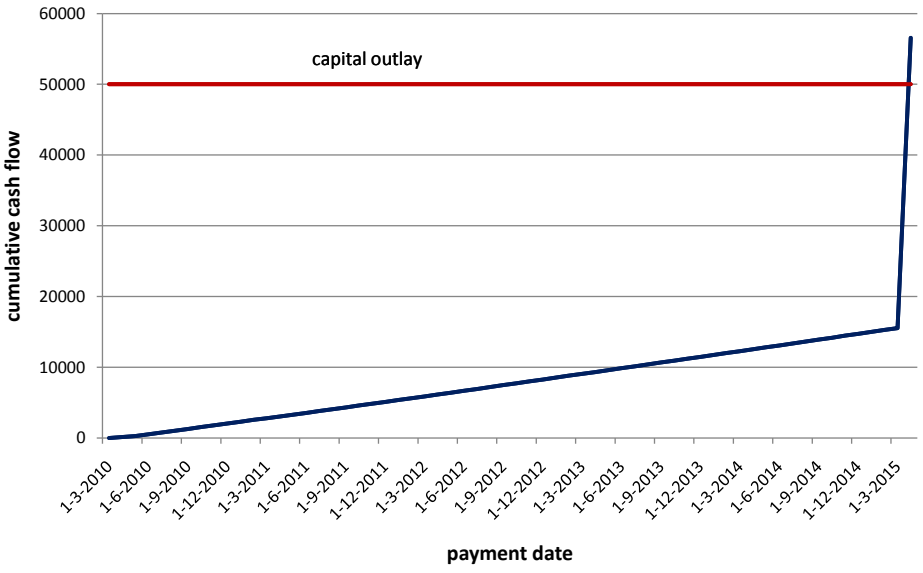


Figure 8.7: Cumulative discounted cash flows to a note in tranche E

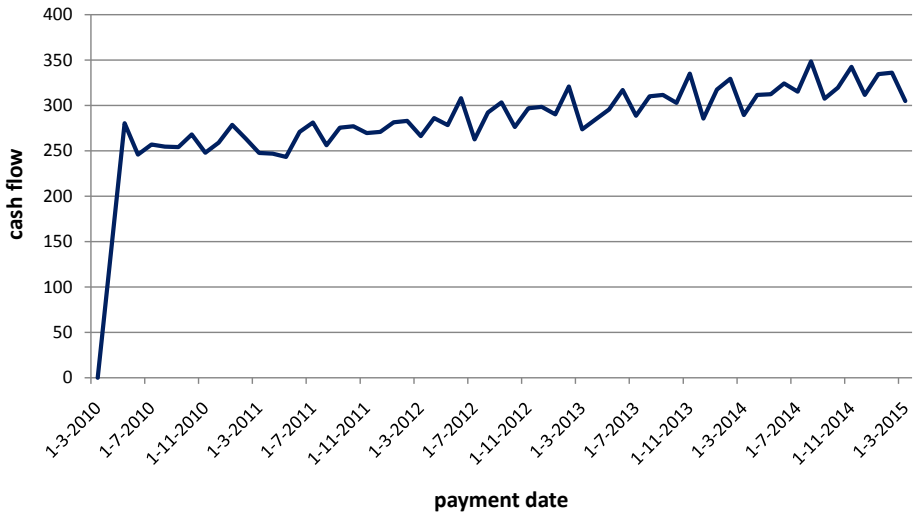


Figure 8.8: Monthly cash flows to a note in tranche E

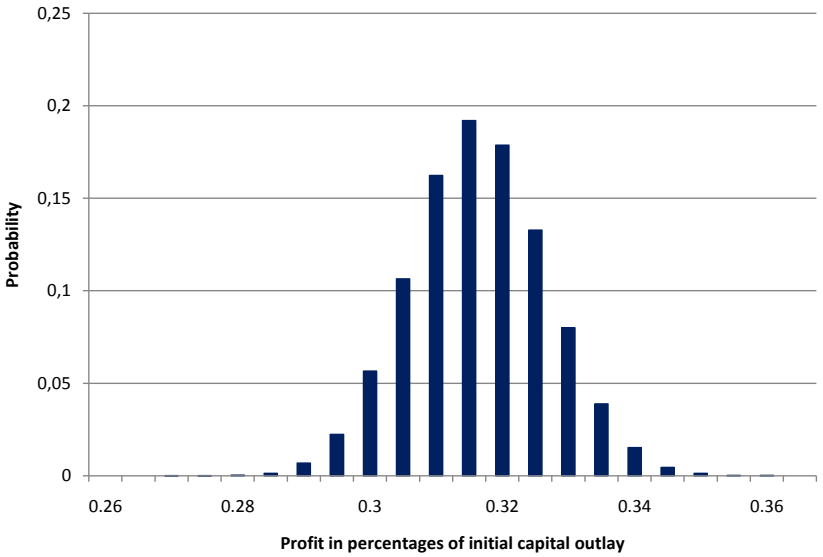


Figure 8.9: Probability distribution of profit for tranche A1 discounted at 90 bps over 1M Euribor

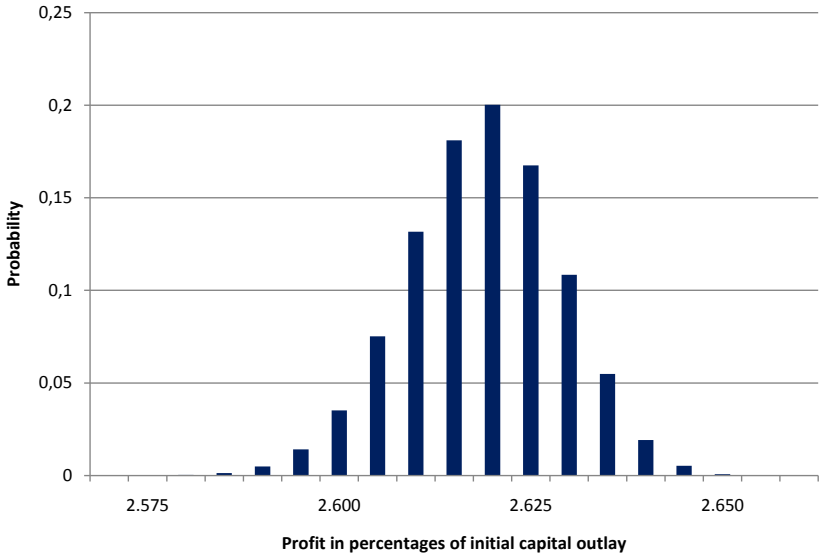


Figure 8.10: Probability distribution of profit for tranche A2 discounted at 90 bps over 1M Euribor

We can conclude from this example that the credit risk for DMBS XV is only minor; the reserve account and the excess spread can in all cases cover the incurred losses. In figure 8.11 two different realisations of incurred losses in DMBS XV according to our models are displayed. Note that in our simulation runs there can not be an incurred loss in the first 18 months due to the fact that we assume that a foreclosure process takes this amount of time. In general for structures similar to that of DMBS XV and with a Dutch mortgage pool there is only minor credit risk; we can highlight this fact by a numerical example. Let us assume, as for DMBS XV holds, that the interest swap generates an excess spread of 50 basispoints per year and that payment dates are monthly. If we do not account for missed interest on the mortgages which have defaulted but are not foreclosed yet, than the excess spread can not cover all incurred losses at a certain payment date if incurred losses exceed $\frac{0.5\%}{12} = 0.042\%$ of outstanding principal at that specific month. This is for Dutch mortgages a very high percentage, since default rates are very low and LGD is also low in the Netherlands. Even more, if

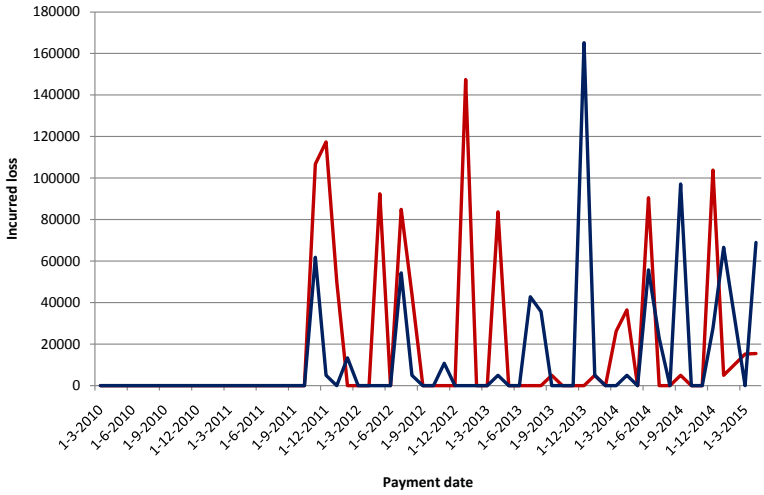


Figure 8.11: Two different realisations of incurred losses for DMBS XV

at a specific payment date an extraordinary high loss would be incurred, the reserve account forms an extra cushion agaisnt losses. So, only if for several payment dates on a row incurred losses are exceptionally high, a loss will be incurred on the most junior notes.

Contrasting, early repayments do cause uncertainty about the value of a note. Although in our example this only holds for the most senior notes in the tranches A1 and A2. The reason that the more junior notes are not affected by early repayments is that early repayments are not high enough in any realisation of our simulation to fully redeem the tranche A2 notes before the FORD. Therefore the notes in tranche B till E always receive only one principal cash flow at the call date of the transaction.

Furthermore, the applied discount spread plays an important role in the determination of the value of a note. This discount spread is a source of uncertainty and ambiguity of the value of a note in the market. If we would lower the applied discount spread of 90 basispoints, the entire probability distribution of the value of a note would shift to higher values. Note that the probability distribution for the notes in tranche B till E consist of only one fixed value with probability 1.

Chapter 9

Conclusions and further research

In this final chapter we will summarize the steps that were taken in this project and the conclusions that we were able to draw from it, in the first section. The second section discusses further research to be done to improve the RMBS pricing tool that we developed in this project.

9.1 Conclusions

In this paper we have presented a method for modelling the distribution of the value of RMBS notes based on individual data of the underlying mortgage pool. The motivation for this research was the fact that regulatory supervisors have, as a result of the credit crisis, requested more transparency from issuers of RMBS notes. Investors have therefore at their disposal, in the near future, loan-level data on the underlying mortgage pool of an RMBS transaction. NIBC is an issuer as well as an investor in RMBS notes and is therefore confronted with the question how to purposefully employ the available data to arbitrage free value an RMBS note. This has been the starting point of the research performed in this thesis.
The valuation of an RMBS note has a stochastic part, the cash flows from the

mortgage pool, and a deterministic part, the allocation of these cash flows to the different notes determined by the transaction structure. Besides interest payments, there are two sources influencing the timing and amount of principal cash flows from a mortgage: default and early repayment. We started the research by a general overview on the available literature on modelling these two processes, see chapter 4. We argued to apply survival analysis and specifically a Cox proportional hazards model. There are several reasons for this particular choice, namely the Cox model:

- enables us to model the default and early repayment intensity over the lifetime of a mortgage, which we expect not to be constant.
- is capable of coping with censored observations. A censored observation is a observation for which no exact event time is known.
- offers the possibility to explicitly model the competing risk of terminating a mortgage by either default or early repayment.

We discussed the characteristics of the Cox proportional hazards model in chapter 5. In this chapter we also described the different approaches, as an extension to the Cox model, to explicitly model the competing risk setting and we selected the cause-specific hazard approach by Kalbfleish and Prentice. The value of the parameters in the model by Cox can be estimated by the non-parametric partial likelihood approach. In chapter 6 we have described this method as well as methods to deal with ties in the data and the fact that our study is a delayed entry study, meaning that some mortgages are only observed a few years after they have been issued.
Chapter 7 described the characteristics of the data set we had at our disposal for estimating the Cox model for default and early repayment. This chapter also gave an outline of the model development steps.
Chapter 8 described the results for the probability of default and the probability of early repayment model. We have shown that the probability of default for a mortgage is higher if:

- the ratio of loan to foreclosure value is higher;
- the borrower has a registered negative credit history;
- the ratio of main income to total income associated with the loan is higher;
- there is only one registered borrower;
- the income of the borrower is not disclosed to the lender, but to an intermediary.

For early repayment of a mortgage, we have shown that the probability of occurrence for a mortgage is higher if:

- the ratio of loan to foreclosure value is higher;
- the (youngest) applicant is younger;
- the total income of the borrower(s) is lower;
- the 3-months Euribor is higher;
- it is an interest reset date;
- the refinancing incentive is higher.

In this last chapter we also applied the models for default and early repayment to forecast cash flows for an RMBS transaction. For this purpose we have developed in Delphi ® (Embarcadero; San Francisco, CA, USA) integrated development environment a user-friendly tool to value a note of an RMBS transaction. The tool offers the user the possibility to select the input data and specify the characteristics of the specific RMBS. We have applied the model to a transaction issued by NIBC, called DMBS XV, and the results are displayed in chapter 8. We concluded for this transaction that credit risk is only minor and therefore defaults have almost no influence on the value of a note in DMBS XV. However, early repayments and the discount spread

applied in discounting cash flows influence the value of a note. The discount spread depends on the funding cost and may be different for each investor. For tranches that might be (partially) redeemed before the FORD we obtain a probability distribution of the value of a note depending on the timing of early repayments. This distribution shifts with the choice of the discount spread. Tranches that are not redeemed before the FORD in any realisation of the simulation process (and also never incur a loss) have only one possible value-outcome, which also shifts with the choice of the discount spread.
The developed model can be used for the analysis of any Dutch RMBS as long as loan-level data on the underlying mortgage pool is available, which by regulation will be the case for any newly issued transaction. Also the structure should fit within the options the tool offers. However, the model cannot be easily adopted for other countries. The reason is that other countries might have a completely different mortgage market and it would be unrealistic to assume that the model can be extended to these markets without any changes. Nevertheless, the methods described in this paper can be adopted for other countries, provided that enough historical data on residential mortgages originated in that market is available.

9.2 Further research

With this project we have made an important step towards a user-friendly RMBS pricing tool based on loan-level data. However, since we build this tool from scratch, we had to make some simplifying assumptions. In this section we describe which are the main areas for further research to improve the model.

1. When we started this project, the expectation was that we would find one or more market related parameters which influence the probability of default. The idea was to generate a large number of possible realisations of these parameters, which would give the same number of different realisations of the probabilities of default. This stochastic

element completely fell out the analysis, since we did not discover such a variable and the probabilities of default became static. With the framework for the valuation tool standing, it would be very relevant to do further research on possible market parameters influencing the probability of default. In this way it would also be possible to stress the market and analyse the effect on the NPV of the RMBS notes.

2. We use for LGD a simple formula expressing the loss for a defaulted mortgage as a fixed amount plus an amount depending on the loan-to-foreclosure value (LTFV). However, it would be interesting to model the LGD in a more sophisticated manner. For example the liquidity in the market could be a good indicator of the actual proceeds from a foreclosure process. From the credit crisis it became evident that, even though a house is in good shape, it might be unsaleable due to an illiquid market. Also, by relating the LGD to the value of the property at the last taxation, we do not account for the up-to-date market value of the property. We could incorporate an estimator for the market value of the property by using the house price index. Since house prices are highly autocorrelated, prediction of a price index is complicated and we decided that it was outside the scope of this research. Hypothetically, making the LGD a stochastic variable relating to market conditions would further improve the model. This could also bring to light the dependency we expect between PD and LGD. Supposedly, when mortgage default rates are higher (a higher PD) also the incurred loss (LGD) is higher, possibly because a higher PD results in a less liquid market which in turn results in a higher LGD.

3. We have made some simplifying assumptions related to the recovery of a defaulted mortgage. We assumed that the probability of recovery is for each mortgage the same and we modelled it as a constant which was derived from the data. Also, according to the model a mortgage which has recovered from a default becomes an ordinary performing

mortgage again. This is, however, not very realistic since defaulted mortgages have evidently a higher probability of going in default again. A first improvement would therefore be to include an extra variable in modelling the probability of default, which is an indicator of the event of previous default. A second, more complicated step, would be to model the probability of recovery based on the mortgage characteristics.

4. The available data was quite restricted; it spans only 6.5 years (including the credit crisis, which is not a very representative period) and includes no more than 1,760 defaults. A database that spans over a longer period could improve the accuracy of the model and possibly bring to light more explanatory variables.

5. Further research has to be done in the direction of handling missing variable values. We have chosen to replace all missing values by the median of the observed values, however in literature other methods are described as well, see for an extensive overview Schafer and Graham (2002). It was outside the scope of this research to further investigate the effect that different methods have on the estimated model and the sensitivity and significance of the parameters.

6. In our search through literature we found a lot of articles related to the appropriateness of the model assumptions and assessment of the overall goodness-of-fit for a Cox proportional hazards model. In spite of this, none of these methods were easily extendable or even appropriate for a model with time-varying covariates. It would therefore be a useful extension to literature to do research on easy to implement methods to analysis the goodness-of-fit for a Cox model with time-varying covariates.

7. As discussed in the previous section we think that the developed model is not easily applicable to other countries. Nevertheless, a similar research could be done on historical data of residential mortgages from

another country and then it would require only minor adjustments to adopt the model to include RMBS transactions from this country.

8. The RMBS tool offers the user some options to specify the structure of a specific transaction. The transactions we have studied, all issued by NIBC, can be matched with these options. However, more flexibility to also match the specifics of other transactions, especially related to triggers in the principal and interest waterfall, would broaden the applicability of the tool. For this purpose it would be useful to do a dedicated research on the structures of RMBS transactions available in the market.

Bibliography

[1] Alink, B. (2002), Mortgage prepayments in the Netherlands. *Enschede University*, PhD Thesis.

[2] Andersen, P. and Gill, R. (1982), Cox's regression model for counting processes: a large sample study. *The Annals of statistics*, Vol. 10, No. 4, pp. 1100-1120.

[3] Association for Financial Market in Europe, www.afme.eu (2011).

[4] Banasik, J., Crook, J. and Thomas, L. (1999), Not if but when will borrowers default. *The Journal of the Operational Research Society*, Vol. 50, No. 12, pp. 1185-1190.

[5] Breslow, N. (1972), Discussion following "Regression models and life tables" by D.R. Cox. *Journal of the Royal Statistical Society*, ser. B, Vol. 34, No. 2, pp. 187-220.

[6] Breslow, N. (1974), Covariance analysis of censored survival data. *Biometrics*, Vol. 30, No. 1, pp. 89-100.

[7] Burkhard, J. and Giorgi, De, E. (2004), An intensity based non-parametric default model for residential mortgage portfolios. *Risk lab report*, available at http://www.risklab.ch.

[8] Caflisch, R. (1998), Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, Vol. 7, pp. 1-49.

[9] Campbell, T. and Dietrich, J. (1983), The determinants of default on insured conventional residential loans. *Journal of Finance*, Vol. 38, No. 5, pp. 1569-1581.

[10] Cox, D. (1972), Regression models and life-tables. *Journal of the Royal Statistical Society*, Ser. B, Vol. 34, No. 2, pp. 187-220.

[11] Efron, B. (1977), The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, Vol. 72, No. 359, pp. 557-565.

[12] Epperson, J., Kau, J. Keenan, D. and Muller, W. (1985), Pricing default risk in mortgages. *Real estate economics*, Vol. 13, No. 3, pp. 261-272.

[13] Fine, J. and Gray, R. (1999), A proportional hazards model for sub-distribution of a competing risk. *Journal of the American Statistical Association*, Vol. 94, No. 446, pp. 496-509.

[14] Fisher, L. and Lin, Y. (1999), Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, Vol. 20, pp. 145-157.

[15] Foster, C. and Van Order, R. (1985), FHA terminations: A prelude to rational mortgage pricing. *Real estate economics*, Vol. 13, No. 3, pp. 273-291.

[16] Guo, S. (2010), Survival analysis. *Oxford University Press.*

[17] Gray, R. (1988), A class of k-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics*, Vol. 16, No. 3, pp. 1141-1154.

[18] Hendershott and van Order (1987), Pricing mortgages: an interpretation of the models and results. *Journal of financial services research*, Vol. 1, No. 1, pp. 19-55.

[19] Hertz-Picciotto, I. and Rockhill, B. (1997), Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, Vol. 53, No. 3, pp. 1151-1156.

[20] Hosmer, D, Lemeshow, S. and May, S. (2008), Applied survival analysis: regression modeling of time-to-event data. *John Wiley and Sons*, New York.

[21] Jackson, J. and Kasserman, D. (1980), Default risk on home mortgage loans: a test of competing hypotheses. *Journal of risk and insurance*, Vol. 47, No. 4, pp. 678-690.

[22] Kalbfleisch, J. and Prentice, R. (1980), The statistical analysis of failure time data. *Wiley*, New York.

[23] Kaplan, E. and Meier, P. (1958), Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Vol. 53, No. 282, pp. 457-481.

[24] Kau, J. and Slawson, V. (2002), Frictions, heterogeneity and optimality in mortgage modeling. *Journal of Real Estate Finance and Economics*, Vol. 24, No. 3, pp. 239-260.

[25] Kimball, A. (1969), Models for the estimation of competing risks from grouped data. *Biometrics*, Vol. 25, No. 2, pp. 329-337.

[26] Klein, J. and Bajorunaite, R. (2004), Inference for competing risks. *Handbook of Statistics*, Vol. 23, pp. 291-311.

[27] Klein, J. and Wu, R. (2004), Discretizing a continuous covariate in survival studies. *Handbook of Statistics*, Vol. 23, pp. 27-42.

[28] Krystul, J. (2006), Modeling of stochastic hybrid systems with applications to accident risk assesment, PhD thesis. *Twente University*.

[29] Lamarca, R., Alonso, J., Gomez, G. and Muñoz, A. (1998) Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population. *Journal of Gerontology*, Vol. 53a, No. 5, pp. 337-343.

[30] Latouche, A., Porcher, R. and Chevret, S. (2005), A note on including time-dependent covariate in regression model for competing risks data. *Biometrical Journal*, Vol. 47, No. 6, pp. 807-814.

[31] Li, J. (2010), Cox model analysis with the dependently left truncated data, Master thesis. *Georgia State University*.

[32] McDonald, R, Matuszyk, A. and Thomas, L. (2010), Application of survival analysis to cahs flow modelling for mortgage products. *OR insight*, Vol. 23, pp. 1-14.

[33] Merton, R. (1974), On the pricing of corporate debt: the risk structure of interest rates. *Journal of finance*, Vol. 29, No. 2, pp. 449-470.

[34] NIBC Bank N.V.(2005), Dutch MBS XII B.V. Prospectus. *internal document*.

[35] NIBC Bank N.V.(2010), Dutch MBS XV B.V. Prospectus. *internal document*.

[36] Prentice, R., Kalbfleisch, J., Peterson, A. Florunoy, N., Farewell, T. and Breslow, N. (1978), The analysis of failure times in the presence of competing risks. *Biometrics*, Vol. 34, No. 4, pp. 541-554.

[37] Putter, H. Fiocco, M. and Geskus, R. (2007), Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, Vol. 26, No. 11, pp. 2389-2430.

[38] Quercia, R. and Stegman, M. (1992), Residential mortgage default: a review of the literature. *Journal of housing research*, Vol. 3, No. 2, pp. 341-379.

[39] R Development Core Team (2011), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[40] Royston, P. and Altman. D. (1994), Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling (with discussion). *Applied Statistics*, Vol 43, No. 3. pp. 429-467.

[41] Royston, P. and Sauerbrei, W. (2008), Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modeling continuous variables. *Wiley Series in Probability and Statistics*.

[42] Satagopan, J. Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. and Auerbach, A. (2004), A note on competing risks in survival data analysis. *British Journal of Cancer*, Vol. 91, No. 7, pp. 1229-1235.

[43] Schafer, J. and Graham, J. (2002), Missing data: our view of the state of the art. *Phychological Methods*, Vol. 7, No. 2, pp. 147-177.

[44] Tai, B., Machin, D., White, I. and Gebski, V. (2001), Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. *Statistics in medicine*, Vol. 20, No. 5, pp. 661-684.

[45] Webb, B. (1982), Borrower risk under alternative mortgage instruments. *The Journal of Finance*, Vol. 37, No. 1, pp. 169-183.

[46] Weng, Y. (2007), Baseline survival function estimators under proportional hazards assumption. *Institute of Statistics, national University of Kaohsiung*.

[47] Wong, J., Fung, L. Fong, T. and Sze, A. (2004), Residential mortgage default risk and the loan-to-value ratio. *Hong Kong monetary authority quarterly bulletin, December 2004*.

Appendix A

Derivation likelihood function

Let us suppose that n subjects give rise to the data $(\tau_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, where τ_i is the observed survival time, δ_i is the censoring indicator ($\delta_i = 0$ if the i-th subject is censored and 1 if an event happened) and $\mathbf{X}_i$ is the covariate vector of the i-th subject. In simplest terms, the likelihood function is an expression that yields a quantity similar to the probability of occurrence of the observed data under the model.
We construct the actual likelihood function by considering the contribution of subjects for which an event is measured separately from the contribution of censored subjects. In case of the triplet $(\tau_i, 1, \mathbf{X}_i)$ we know the exact survival time of subject i to equal τ_i. It's probability to have this exact survival time is given by $f(\tau_i|\mathbf{X})$ which is defined analogous to definition (5.1). For the triplet $(\tau_i, 0, \mathbf{X}_i)$ we know that the survival time of subject i was at least τ_i, this probability equals $S(\tau_i|\mathbf{X})$. Furthermore we assume that the censoring time of subject i is a random variable with survival and density function $G(t|\mathbf{X}_i)$ and $g(t|\mathbf{X}_i)$ respectively. In general we can write the contribution of observation i to the likelihood as

$$\left[f(\tau_i|\mathbf{X}_i)G(\tau_i|\mathbf{X}_i)\right]^{\delta_i}\left[S(\tau_i|\mathbf{X}_i)g(\tau_i|\mathbf{X}_i)\right]^{1-\delta_i}\ . \tag{A.1}$$

Since the censoring time is non-informative we can rewrite (A.1) as

$$[f(\tau_i|\mathbf{X}_i)]^{\delta_i} [S(\tau_i|\mathbf{X}_i)]^{1-\delta_i} . \tag{A.2}$$

As the observations are assumed to be independent, the likelihood function is the product of the expression in (A.2) over the entire sample. We can now rewrite the likelihood function, of which the only unknown parameters are the vector β and the baseline function $\lambda_0(t)$, as

$$L(\boldsymbol{\beta}, \lambda_0(t)) = \prod_{i=1}^{n} \left([f(\tau_i|\mathbf{X}_i)]^{\delta_i} [S(\tau_i|\mathbf{X}_i)]^{1-\delta_i} \right) . \tag{A.3}$$

We can use the relation $\lambda(t|\mathbf{X}) = f(t|\mathbf{X})/S(t|\mathbf{X})$, see equation (5.6), to rewrite (A.3) to

$$L(\boldsymbol{\beta}, \lambda_0(t)) = \prod_{i=1}^{n} \lambda(\tau_i|\mathbf{X}_i)^{\delta_i} S(\tau_i|\mathbf{X}_i) . \tag{A.4}$$

From section 5.1 it holds that $S(t|X) = \exp\left(-\int_0^t \lambda(u|\mathbf{X})du\right)$ and for the Cox model we have $\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp\left(\boldsymbol{\beta}^T\mathbf{X}\right)$. We can therefore express the likelihood function for the Cox model as

$$L(\boldsymbol{\beta}, \lambda_0(t)) = \prod_{i=1}^{n} \left[\lambda_0(\tau_i) \exp\left(\boldsymbol{\beta}^T\mathbf{X}_i\right)\right]^{\delta_i} \exp\left[-\int_0^{\tau_i} \lambda_0(u) \exp(\boldsymbol{\beta}^T\mathbf{X}_i)du\right] . \tag{A.5}$$

The corresponding log-likelihood equation is

$$l(\boldsymbol{\beta}, \lambda_0(t)) = \sum_{i=1}^{n} \left(\delta_i \left[\ln(\lambda_0(\tau_i)) + \boldsymbol{\beta}^T\mathbf{X}_i\right] - \exp\left(\boldsymbol{\beta}^T\mathbf{X}_i\right) \int_0^{\tau_i} \lambda_0(u)du \right) . \tag{A.6}$$

This full likelihood function contains an unspecified baseline hazard function so that the estimate of $\boldsymbol{\beta}$ is difficult to obtain. Cox developed for this purpose the partial likelihood method described in section 6.1. The full likelihood function is used to estimate the baseline after obtaining the estimate for $\boldsymbol{\beta}$ as described in section 6.2.

Appendix B

Model fitting

In this appendix we give a detailed overview of the steps we took to fit a survival model for the default and early repayment of a mortgage; these steps have been discussed in section 7.2. The purpose of this process is to find the relevant variables with the corresponding $\boldsymbol{\beta}$ vector. To this end we maximize the likelihood function (6.12) over the vector $\boldsymbol{\beta}$. When the final model is obtained we can calculate the baseline by formula (6.13). The variables we have at our disposal are listed in table 7.1. While most of the variables might have an influence on both default and early repayment, other variables will only be part of the initial variables set of one of the two analyses. Examples of the last case are the reset date and the refinancing incentive, which will both not be considered for the probability of default.

The data set contains some missing data for the income and therefore also for LTI, IPTI and SMI. We assume that the probability that a value is missing does not depend on the outcome or on any of the covariates measured, i.e. data is missing completely random. We therefore replace the missing values by the median of the data.

The first section of this appendix describes the steps taken for the default model and the second section those of the early repayment model. The final models are discussed in section 8.1 and 8.2 respectively.

B.1 Default model

step 1-4: multivariable model

We perform a univariable analysis for all variables that could play a role in the default process of a mortgage and we find at the 20% level only income, value of the property, outstanding balance and area of the property to be insignificant. The next step is to fit a multivariable model containing all the remaining variables. However the LTFV is highly correlated to the LTiFV and for this reason we will not use both variables in our final model. Something similar holds for the age of the oldest and the age of the youngest applicant. Therefore we will use four different settings to fit the initial multivariable model by combing LTFV or LTiFV with the age of the oldest applicant or the age of the youngest applicant.

For all four models we first remove the interest and swap rates and observe only small changes in the parameter values. Next we remove the interest on the mortgage, thirdly the LTI, and in the fourth step the age of the applicant. The IPTI is still marginally significant in both models, but the effect is practically zero and removing the IPTI does barely influence the performance of the model or the values of the other covariates. We therefore also remove the IPTI from the model.

The next step is to add, one at the time all variables initially excluded from the multivariable analysis. In this step we obtain that those variables are also not significant in the presence of the other variables and they can therefore be removed again. We are now left with two quite similar models which are summarized in table B.1a and B.1b. The Wald statistic z for the individual β's is calculated as $z = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}$ and it is together with it's two-sided p-value displayed.

To decide which model will be our preliminary main effects model, we compare the performance of both models. From table B.2 we see that the model based on LTFV scores better on all test statistics and therefore we will use this model in the remainder of the analysis.

	Coefficient	exp(coef)	se(coef)	z	P-value
LTFV	0.018573	1.018747	0.000597	31.119	< 2e-16
BKR	1.909917	6.752527	0.091695	20.829	< 2e-16
SMI	0.012634	1.012714	0.001621	7.794	6.55e-15
Number applicants	-0.14526	0.864796	0.055799	-2.603	0.00923
Advisor verified	1.403831	4.070766	0.068946	-20.361	< 2e-16

(a) LTFV

	Coefficient	exp(coef)	se(coef)	z	P-value
LiTFV	0.022971	1.023236	0.000921	24.945	<2e-16
BKR	1.879058	6.547333	0.091934	20.965	<2e-16
SMI	0.014715	1.014824	0.001641	8.965	< 2e-16
Number applicants	-0.09892	0.905818	0.036019	-2.746	0.00603
Advisor verified	1.452564	4.274059	0.070723	20.539	< 2e-16

(b) LTiFV

Table B.1: Preliminary models for PD with LTiFV and LTFV

step 5: scale continuous covariates

The next step is to check the scale of the continuous covariates, in our case the LTFV and the SMI; a summary of these variables is given in table B.3. For the SMI almost 75% of the data has a value of 100, this is partly due to the fact that we changed the missing values for SMI to 100. For about 10% of the mortgages no income details are registered, and consequently for those mortgages we have no information on SMI either.

To have a first impression of the scale of the continuous covariates we will apply the quartile design variable method. Since the data for SMI is so poorly distributed we will only apply this method to the LTFV, which is displayed in figure B.1. To ensure that the graph is not too much disturbed by heavy outliers, we did not take into account the 0.1% biggest outliers when determining the midpoint of the last quartile.

It is difficult to tell from figure B.1 whether the plot for LTFV indicates a sig-

	Model LTFV	Model LTiFV
LR	1090	1074
Wald	641.7	632.8
Score	369.2	348.6

Table B.2: Comparing model performance

	min	1st Q	Median	mean	3rd Q	max	NA's
LTFV	0	66.94	89	87.94	118	532.5	2471
SMI	0	67	100	85.62	100	100	0

Table B.3: Summary LTFV and SMI data

nificant departure from linearity or is due to a random variation. Therefore we will apply the fractional polynomials method to suggest a transformation of the covariate, see for details on this method section 7.2.3. From the summary of the application of this method in table B.4a, we can conclude that we will apply a transformation of LTFV by taking the square root of the LTFV. Although this transformation gives a better description of the data, it will make the model harder to interpret.

As we can see from table B.4b no transformation of the SMI covariate gives a significant improvement of the model and therefore we keep this covariate linear. We now have the model as in table B.5

step 6-7: interaction terms and overall goodness-of-fit

The final step in the variable selection procedure is to determine whether interaction terms are needed in the model. The only plausible interaction term in our point of view would be the interaction between BKR and Advisor verified, since both have a very large increasing effect on the probability of default thereby possibly overestimating the PD of someone with a BKR registration and an Advisor verified loan. We conclude from the distribution

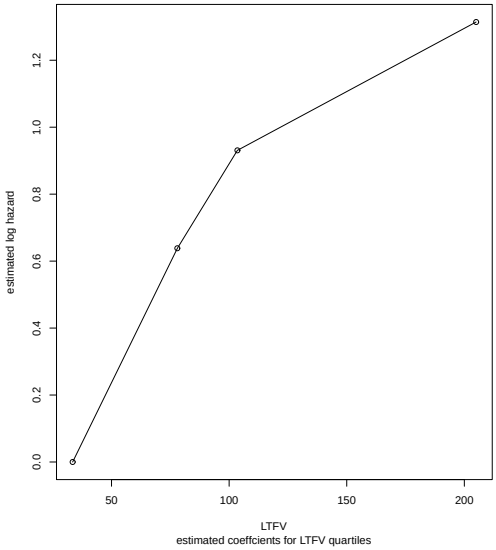


Figure B.1: Graph of estimated coefficients versus quartile midpoints for LTFV

of the data over these variables, see table B.6, that it would not be meaningful to include such an interaction term.

Before proceeding to the final step of checking for the appropriateness of the model, we first want to approach the selected covariates with some common sense. That the selected covariates have an effect on the probability of default is not unexpected and also the size of the effects could be explained, see for more details section 8.1. For the continuous covariates we have checked the scale and made a transformation to one of them. We will also have a look at the scale and definition of the other covariates. The covariates "BKR" and "Advisor verified" are binary variables and therefore there is nothing to check about there scale. The covariate "Number Applicants" can have the values 1 to 10, but only 14,800 observations of the 3,350,022 have more than 2 applicants registered of which only 24 defaults happen. Since the model assumes that an increase of 1 applicant gives the same effect on the probability of default when going from 1 to 2 applicants as for any other increase of 1 applicant, this might not be the most appropriate definition of

	Log likelihood	G for model vs Linear	Approx. p-value	Powers
Not in model	-17218.94			
Linear	-16960.91	0.000	0.000[1]	1
J=1	-16923.27	75.28	0.0000[2]	0.5
J=2	-16919.4	76.56	0.865[3]	0, 0.5

(a) LTFV

	Log likelihood	G for model vs Linear	Approx. p-value	Powers
Not in model	-16960.91			
Linear	-16930.52	0.000	0.000[1]	1
J=1	-16930.21	0.31	0.865[2]	-0.5
J=2	-16925.82	4.70	0.319[3]	0.5, 3

(b) SMI

[1] compares linear model to model without LTFV/SMI

[2] compares the best J=1 model to model with LTFV/SMI

[3] compares the best J=2 model to the best J=1 model

Table B.4: summary of fractional polynomial method for PD

our covariate. We change the covariate to a binary covariate indicating a '0' when there is only one applicant and a '1' when the number of applicants is two or more. As can be seen in table B.7 this model significantly better describes the data and we will therefore carry on our analysis with this model. Hence our final model is the model in table B.8 in which LTFV is the only remaining time-varying covariate, al other covariates are only recorder at issue date. The LTFV may change due to a change in the foreclosure value of the property or a change in the loan amount. The registered foreclosure value can change when a new taxation report is received by the issuer, for example when a borrower wants to increase his mortgage. The loan amount decreases when a borrower decides to partly redeem his mortgage. Finally,

	Coefficient	exp(coef)	se(coef)	z	P-value
$\sqrt{LTFV}$	0.424136	1.528269	0.019092	22.215	<2e-16
BKR	1.872514	6.504628	0.091945	20.366	<2e-16
SMI	0.014187	1.014288	0.001642	8.640	< 2e-16
Number applicants	-0.107295	0.898261	0.055947	-1.918	0.00499
Advisor verified	1.507872	4.517108	0.073043	20.644	< 2e-16

Table B.5: Model PD

	Total mortgages	Defaulted mortgages
All mortgages	70518	1760
Advisor verified	7901	363
BKR	907	131
Advisor verified and BKR	124	12

Table B.6: defaults among Advisor verified and BKR mortgages

we can proceed to checking for the appropriateness of the model and assess the overall goodness-of-fit. The most important aspect to check for in the Cox proportional hazards model is, as the name reveils, the proportional hazards assumption. This assumption states that the hazard of any subject in the sample is a fixed proportion of the hazard of any other subject and the ratio of the hazard of two subjects is constant over time. However, since our model includes time-varying covariates the proportionality assumption is violated.

Standard back testing approaches for Cox model assume that the model does not include time-varying covariates. Also literature on testing for overall goodness-of-fit when the model contains time-varying covariates is very restricted and we did not succeed to find any method understandable and clear enough to implement in R. To determine whether an estimated Cox model fits the data to an acceptable degree is the likelihood ratio test mentioned in section 7.2. The test statistic for our model is 1122 and it is subject to a chi-

	model Number applicants $\in \{1, \ldots, 10\}$	Model Number applicants binary
LR	1116	1122
Wald	709.2	721.3
Score	352.8	359.3

Table B.7: Comparing model performance

	Coefficient	exp(coef)	se(coef)	z	P-value
$\sqrt{LTFV}$	0.412748	1.510964	0.028906	14.279	<2e-16
BKR	1.872541	6.504805	0.138789	13.492	<2e-16
SMI	0.013006	1.013090	0.002266	5.739	9.51e-09
Number applicants$_{bin}$	-0.185020	0.831087	0.080746	-2.291	0.0219
Advisor verified	1.481644	4.400173	0.094783	15.632	< 2e-16

Table B.8: Final model PD

square distribution with 5 degrees of freedom. We have $P(\chi^2(5) > 1122) \approx 0$, indicating that the model including the covariates is significantly better describing the data than a model without the covariates. Also all individual covariates are significant in the model.

One way to graphically show the performance of a regression model is by plotting realisations versus modelled events. To this end we take time-scale as the observation period, instead of the outstanding months. The reason is that it is an in-sample back test where the model is estimated with the time-scale as outstanding months. If we would do the back test with the same time-scale the baseline would perfectly match and the coefficients are the maximum likelihood estimates. By using this different time-scale for the back test, grouping of the mortgage observations is completely different and the test has some properties of an out-of-sample back test. In figure B.2 the number of realised defaults divided by the number of observations for each month are displayed as well as the expected defaults according to the model.

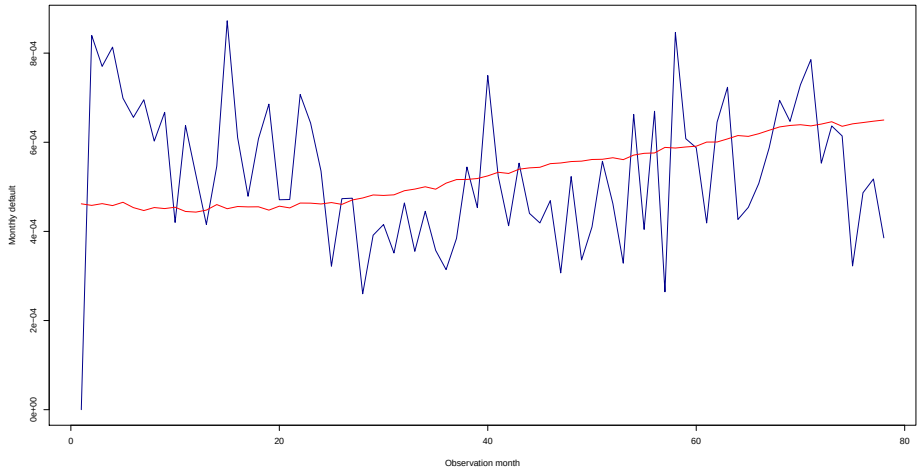


Figure B.2: Graph of realised defaults and modelled defaults by observation month. The blue line indicates the realised defaults divided by the number of observations for each month and the red line displays the expected defaults according to the model divided by the number of observations.

B.2 Early repayment model

step 1-4: multivariable model

Besides the variables listed in table 7.1 we will introduce two new variables related to the interest reset date:

- Two months around the Interest reset date (called IR_{2M}): a binary variable indicating whether an interest reset will take place in 2 months or has taken place less than 2 months ago. So there are five values equal to one around a reset date for this variable.

- Three months around the Interest reset date (called IR_{3M}): equivalent to IR_{2M} but for three months around the reset date.

The logic behind these variable is that borrowers are notified of the new interest rate they will have to pay after the reset date about 2 or 3 months

before the actual reset date and this triggers borrowers to refinance their mortgage in this period. In the same line of reasoning, borrowers tend to forget about refinancing a mortgage in time, consequently they refinance a few months after the actual interest reset date. Even though borrowers do not always make from a economic perspective the most rational decision to refinance at a reset date, which saves them money from not having to pay a prepayment penalty, practice reveals that early repayment is higher in the months around the interest reset date.

In the univariable analysis we found that except for the LTiFV, BKR, LTI, IPTI, SMI, NumbAppl and Reg, all variables were significant at the 20% level. The Refinancing incentive is calculated using the variables interest, 5YR vs 3M swap and Retail spread. Since this variables expresses an incentive to prepay we will not use the related variables in our further analysis. We now proceed with the following variables: LTFV, Advisor verified, Income, Age youngest applicant, Age oldest applicant, Refinancing incentive, Euribor rate, IRDate, IR_{2M} and IR_{3M}. The different variables related to the interest reset date are highly correlated, if a row has a '1' in IRDate it will always also have a '1' in the other two variables, and we therefore cannot use more than one interest reset date related variable in a model. Something similar holds for the age of the youngest and of the oldest applicant. We will therefore use six different settings to fit the initial model, combining the different interest reset date variables with the age of the oldest respectively youngest applicant.
For all six models we remove Advisor Verified as a covariate and the remaining covariates are all significant at the 5% level in all six models. The model with the age of the youngest applicant and the first definition of the interest reset date has the best performance at all test statistics. This model is displayed in table B.9

We expected that an indicator for the months around the interest reset date would give a better description of the data than the exact month of reset. It

	Coefficient	exp(coef)	se(coef)	z	P-value
LTFV	0.001723	1.001725	0.000232	7.441	1,08e-13
Age youngest applicant	-0.008229	0.991805	0.000528	-15.597	<2e-16
Income	-0.006067	0.993952	0.000354	-17.162	<2e-16
Euribor	0.154065	1.166567	0.004118	37.412	<2e-16
resetdate	1.056820	2.877207	0.017608	60.020	<2e-16
Refinancing incentive	1.064326	2.898885	0.020782	51.214	<2e-16

Table B.9: Preliminary model ER

follows that there are significantly more early repayments in these months, especially in one or two months before or after the reset date, as displayed in table B.10.

	Repaid mortgages
At reset date	4732
In 1 or 2 months before and after reset date	944
In 2 months around reset date	5676
In 3th month before and after reset date	268
In 3 months around reset date	5954
Total	33408

Table B.10: Early repayment of mortgages at interest reset date

However, we found that the exact interest reset date was a better predictor for early repayment than the IR_{2M} or IR_{3M} variable. The most plausible explanation for this is that these indicators do not give a higher weight to the exact month of an interest reset date, although from the data it becomes clear that the reset date is the most likely date of early repayment. We therefore define two other variables by changing the '1' for IR_{2M} and for IR_{3M} at the exact date to 2; this gives a kind of stair function. We replace the variable reset date in the model from table B.9 by these variables, but again both models perform worse than the model in table B.9 and therefore

we stick to this model.
Next we add the variables initially excluded from the multivariable analysis one at the time, except LTiFV since it is highly correlated to the LTFV. All variables are found to be still insignificant in the presence of the other covariates and therefore the model in table B.9 will be our preliminary model.

step 5: scale continuous covariates

The next step is to check the scale of the continuous covariates, which are in this case LTFV, Age youngest applicant, Income, Euribor and Refinancing incentive. We first apply the quartile design variables method to see which variables are candidate for a transformation. A summary of these continuous variables is in table B.11 and the graphs of the method for all five continuous covariates are displayed in figure B.3. In the same way as for the PD model, the 0.01% biggest outliers were not taken into account for the drawing of the graphs. We see that for the variables LTFV and Income it is doubful whether the linearity assumption is reasonable and hence we will apply the fractional polynomials method to these variabels.

	min	1st Q	Median	mean	3rd Q	max	NA's
LTFV	0	66.94	89	87.94	118	532.5	2471
Age youngest applicant	-67.00	35.00	41.00	44.09	50.00	110	202
Income (in 1000's)	-11.44	33.35	42.46	47.41	55.07	51840	354658
Euribor	0.635	1.825	2.474	2.718	3.924	5.291	0
Refinancing incentive	0	0.9422	1.0916	1.1303	1.2817	3.0228	2250

Table B.11: Summary continuous covariates in the model

The results of this method are in table B.12. As we can see from table B.12a a transformation for the LTFV of two powers is significantly better than a one power transformation which in turn is significantly better than a linear model. We will therefore proceed with the two-term (0,1) fractional polynomial model.

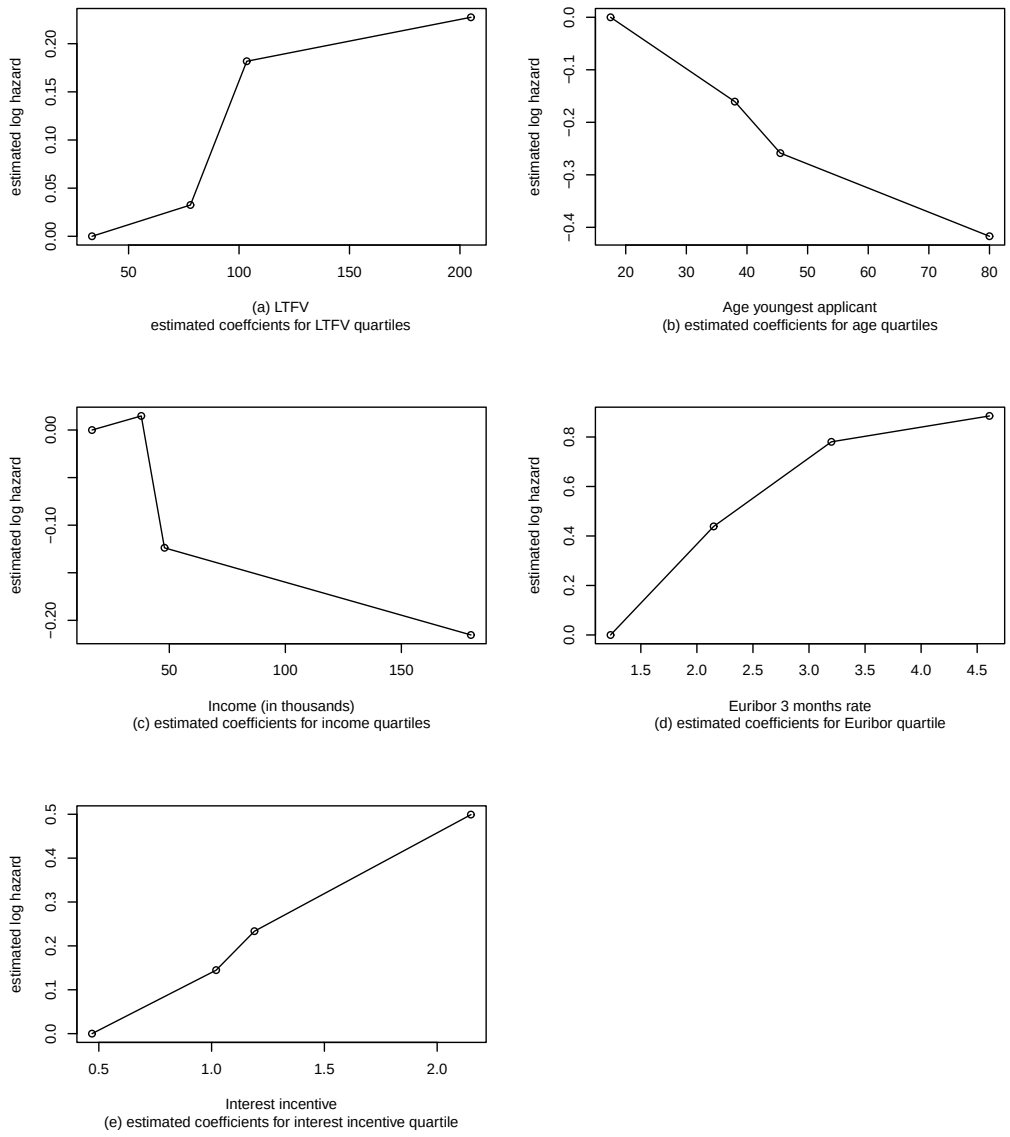


Figure B.3: Graph of estimated coefficients versus quartile midpoints for (a)LTFV, (b)age, (c)Income, (d)Euribor, and (e)refinancing incentive.

For the Income variable table B.12b indicates that a one power transformation is significantly improving the model, while a two-term transformation is not significantly improving the fit of the one power transformation. We will

therefore transform the Income variable by $1/\sqrt{\text{Income}}$.

	Log likelihood	G for model vs Linear	Approx. p-value	Powers
Not in model	-329650.51			
Linear	-329566.16	0.00	0.0000[1]	1
J=1	-329553.54	25.24	0.0000[2]	-1
J=2	-329542.78	21.52	0.0002[3]	0, 1

(a) LTFV

	Log likelihood	G for model vs Linear	Approx. p-value	Powers
Not in model	-329775.21			
Linear	-329557.83	0.00	0.000[1]	1
J=1	-329014.64	1086.38	0.000 [2]	-0.5
J=2	-329011.52	6.24	0.182[3]	-2, 1

(b) Income (in thousands)

[1] compares linear model to model without LTFV/Income

[2] compares the best J=1 model to model with LTFV/Income

[3] compares the best J=2 model to the best J=1 model

Table B.12: summary of fractional polynomial method for ER

This results in our final model as displayed in table B.13, from which it becomes clear that all variables have a significant effect on the probability of early repayment.

step 6-7: interaction terms and overall goodness-of-fit

There are no interaction terms which we expect from the study's perspective to be of interest and we therefore do not include any interaction term and stick with the model in table B.13.

Before proceeding to the final step of checking for the appropriateness of the model, we should first ensure ourselves that the model is intuitively correct.

	Coefficient	exp(coef)	se(coef)	z	P-value
ln(LTFV)	-0.573927	0.563309	0.034304	-16.731	<2e-16
LTFV	0,010762	1.010820	0.000531	20.270	<2e-16
Age youngest applicant	-0.008309	0.991726	0.000543	-15.313	<2e-16
$1/\sqrt{\text{Income}}$	1.858693	6.415348	0.069052	26.917	<2e-16
Euribor	0.154317	1.166860	0.004072	37.897	<2e-16
resetdate	1.009755	2.744928	0.018534	54.483	<2e-16
Refinancing incentive	1.044252	2.841272	0.021116	49.453	<2e-16

Table B.13: Final model ER

A detailed discussion on the theoretical justification of the covariates is in section 8.2. Here we just state that the model is feasible.
In contrast to the PD model which had a lot of binary or integer covariates, the early repayment model mainly has continuous covariates. The scale of these covariates have been checked already. The only remaining covariate is the interest reset date, which is binary and there was some discussion on how to define this variable at the beginning of this section.
A discussion on the difficulties related to assessing the overall goodness-of-fit of a Cox model with time-varying covariates is placed in the previous section. A method to determine whether an estimated Cox model fits the data to an acceptable degree is the likelihood ratio test mentioned in section 7.2. The test statistic for our model is 7549 and it is subject to a chi-square distribution with 7 degrees of freedom. We have $P(\chi^2(7) > 7549) \approx 0$, indicating that the model including the covariates is significantly better describing the data than a model without the covariates. Also all individual covariates are significant in the model.
Furthermore figure B.4 graphically displays the realised and expected early repayments by observation period. For each month between July 2004 and December 2010 the realised early repayments divided by the number of observations in the data are displayed by the blue line. The red line gives the expected early repayments according to the model divided by the number

of observations. For this back test we use observation period as the time scale, where age was used as time scale in fitting the model. This gives the back test, even though it is an in-sample back test, also something of an out-of-sample back test.

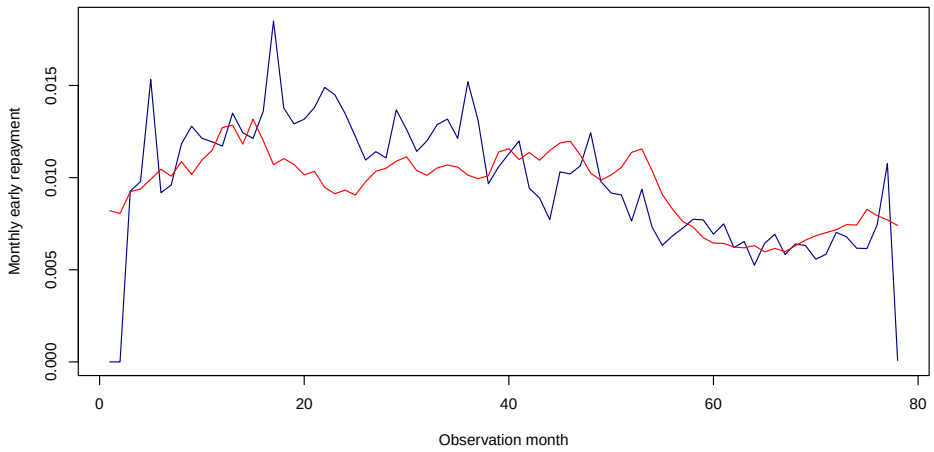


Figure B.4: Graph of realised early repayments and modelled early repayments by observation month. The blue line indicates the realised early defaults divided by the number of observations for each month and the red line displays the expected early defaults according to the model divided by the number of observations.

Appendix C

RMBS valuation tool

The RMBS valuation tool which we developed in this project offers the user several options to match the characteristics of the transaction he intends to value, these include:

- the underlying mortgage pool (the user should select the file in which this data is contained);
- monthly or quarterly payment dates;
- the number of tranches and senior tranches;
- for each tranche, the size in Euro's and the margin over 1-months Euribor in case of monthly payment dates or over 3-months Euribor in case of quarterly payment dates;
- denomination of notes, these must be equal for all tranches;
- whether or not there is initially a reserve account, and if so if it is funded by cash or by the underlying mortgage pool;
- the target size of the reserve account (can also have a value if there is no initial reserve account, the reserve account is than build up from excess spread);

- whether or not the swap agreement guarantees excess spread and if so how much basispoints it is on a yearly basis;
- the issue date, the first payment date and the call date;
- whether the transaction has a clean-up call option, and if so for which percentage of initial outstanding asstets the transaction can be called by the issuer;
- redemption options: pro rata, sequential or pro rata up till a certain tranche;
- triggers related to the principal waterfall;
- day count convention,

and several other transaction specifics. Furthermore, the user should specify the underlying assumptions he would like to make regarding:

- the discount spread applied for valuation;
- the probability of recovery of a mortgage;
- the number of months a foreclosure process takes;
- the parameters regarding the LGD (fixed loss, safety factor, loss factor for NHG, see section 8.3).

Lastly, the user should specify how much simulation runs he would like to perform.

The figures on the next pages give an idea of the tool.

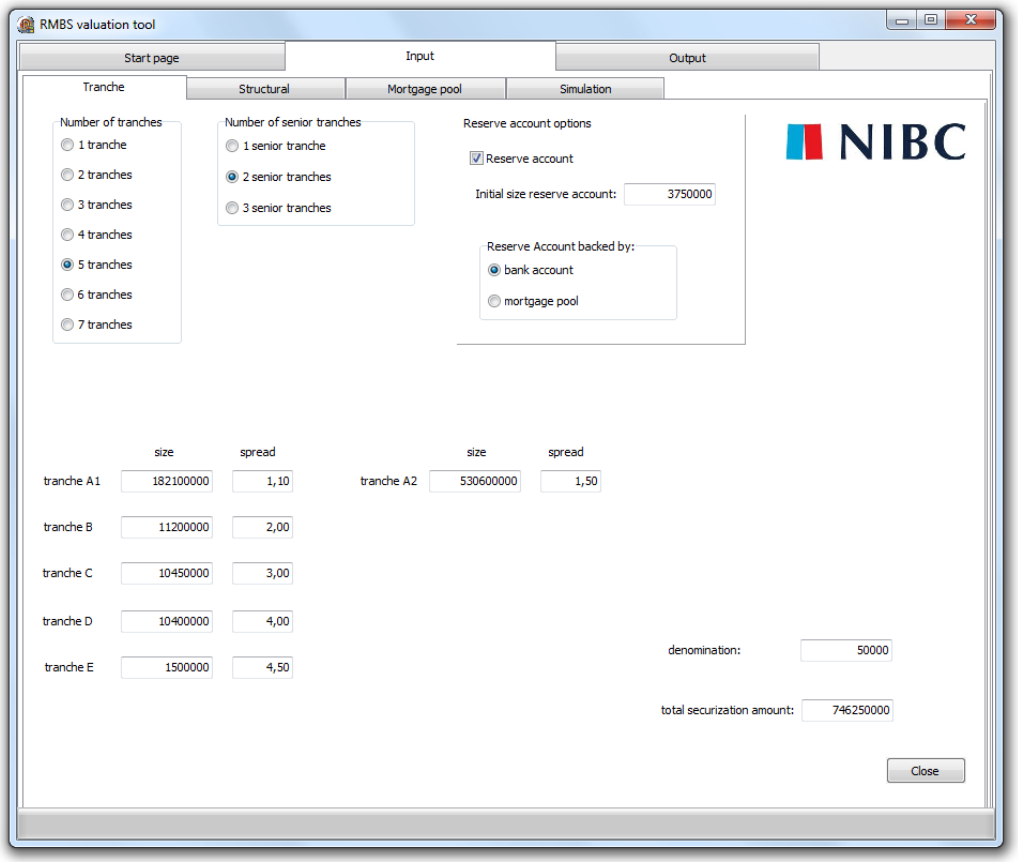


Figure C.1: RMBS valuation tool, tab: tranche input

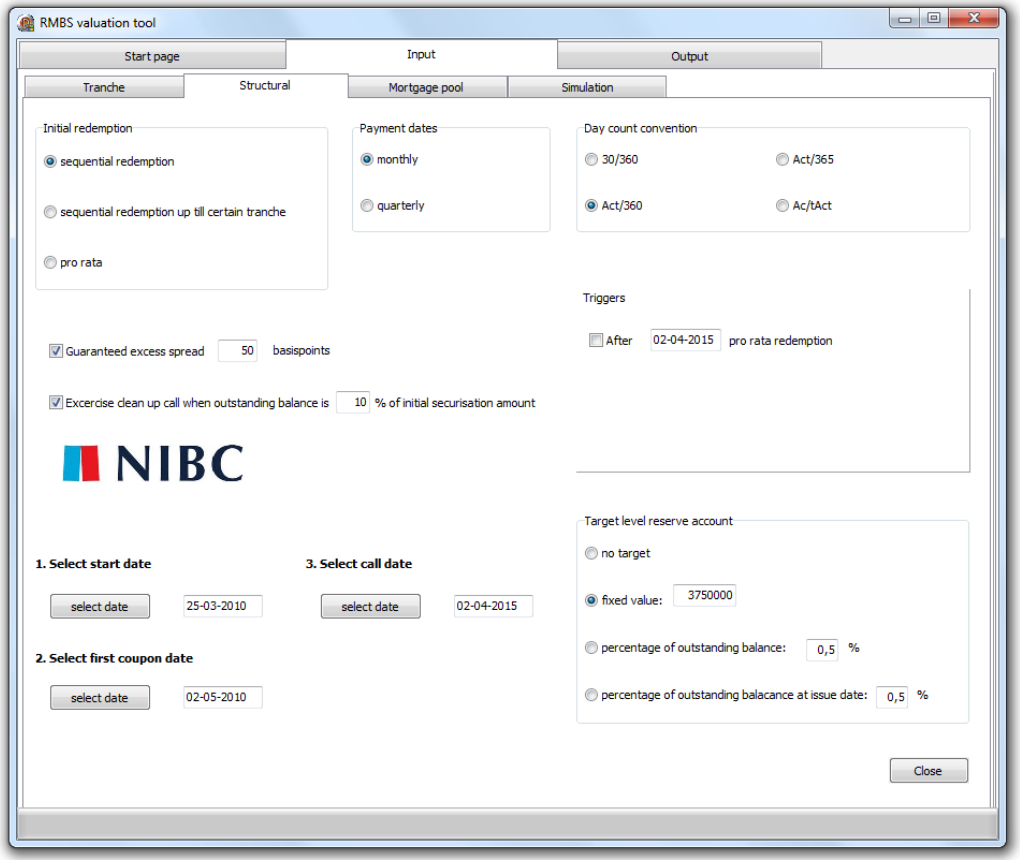


Figure C.2: RMBS valuation tool, tab: structural input

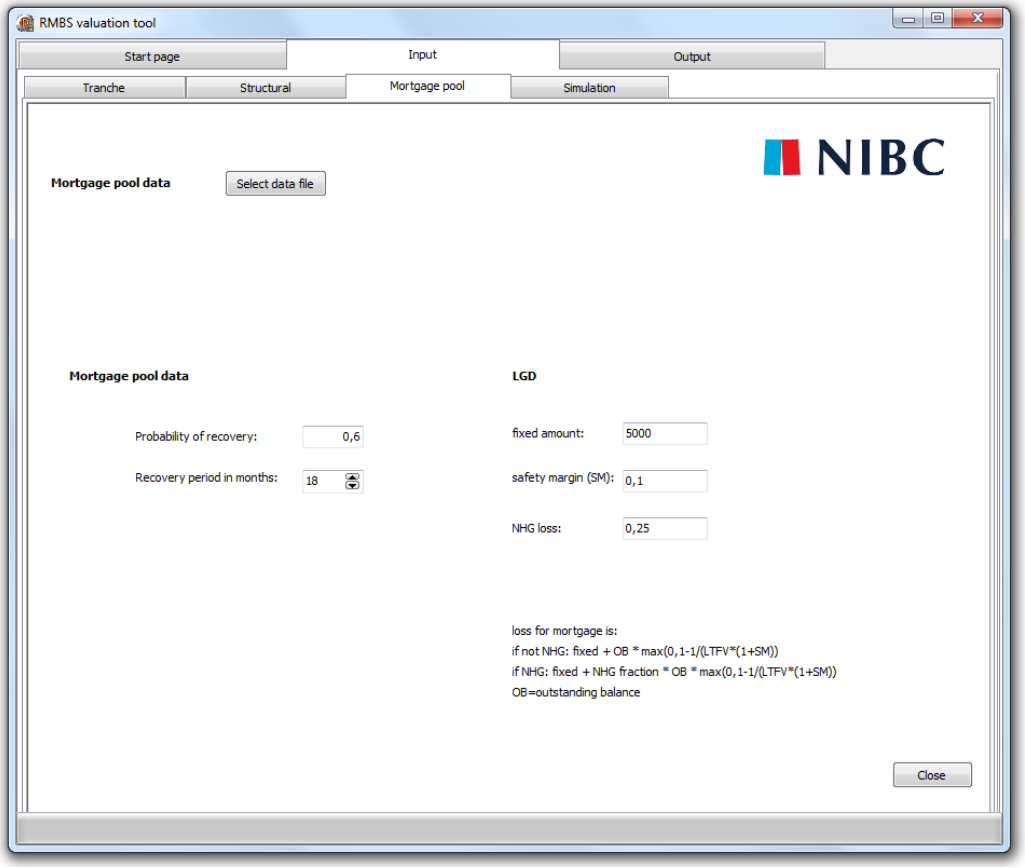


Figure C.3: RMBS valuation tool, tab: mortgage pool input

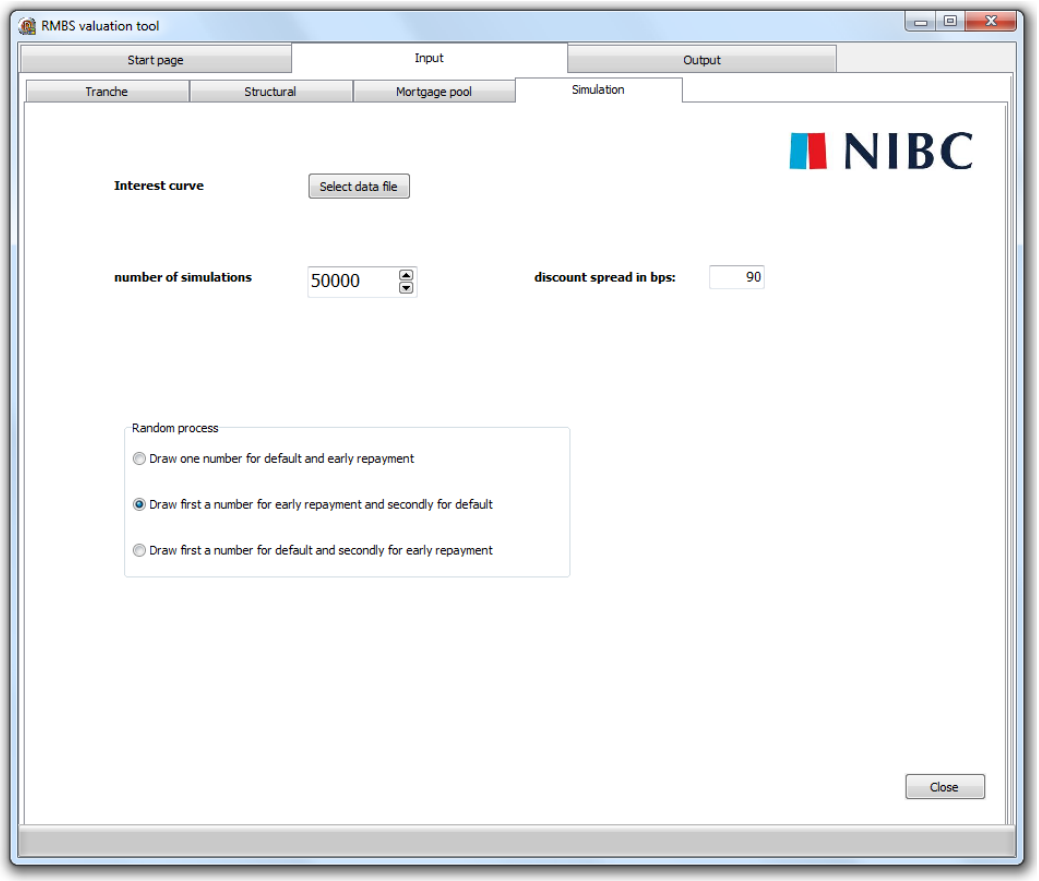


Figure C.4: RMBS valuation tool, tab: simulation input

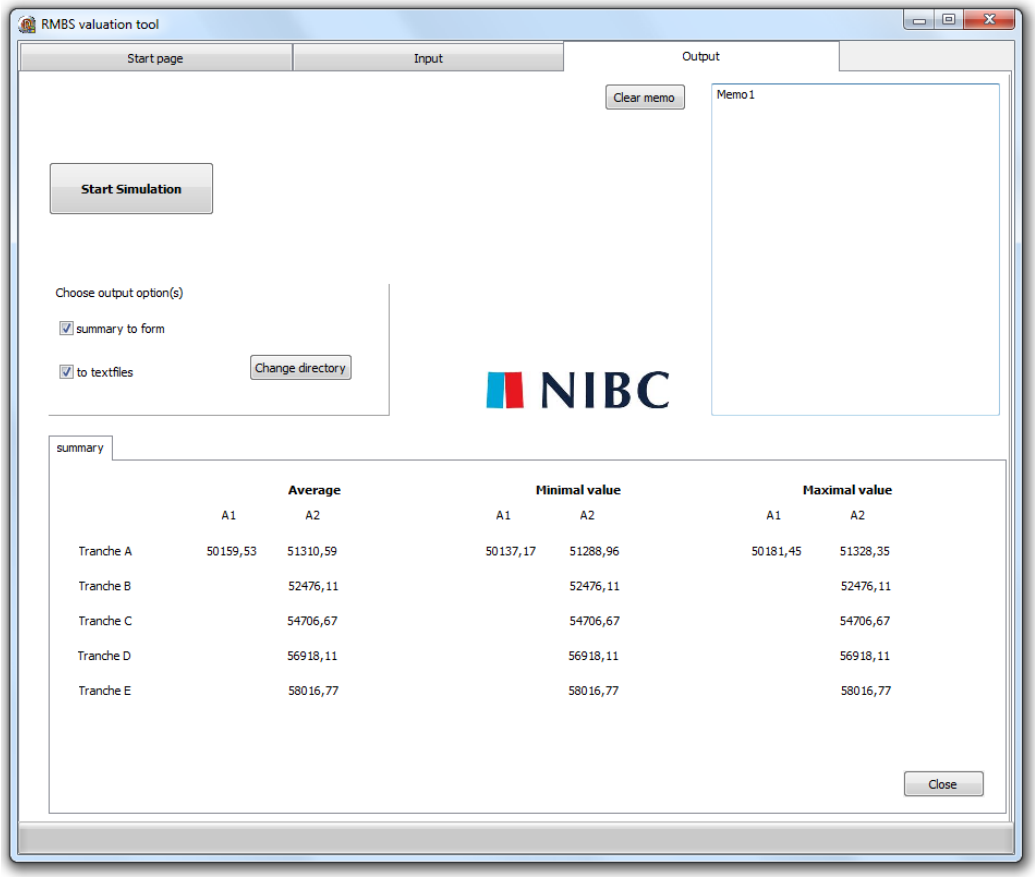


Figure C.5: RMBS valuation tool, tab: output input

Appendix D

Realisations Retail spread

In section 8.4.1 we estimated a model for the retail spread as:

$$RS_t = (1 + Z) \cdot RS_{t-1}$$

Where RS_t is the retail spread at time t and Z is a normal distributed random variable, i.e. $Z \sim N(\mu, \sigma^2)$. We found that $\mu = 0.01766$ and $\sigma = 0.091114$. Figure D.1 shows ten different realisations of the retail spread for the next five years, corresponding to the maturity of DMBS XV.

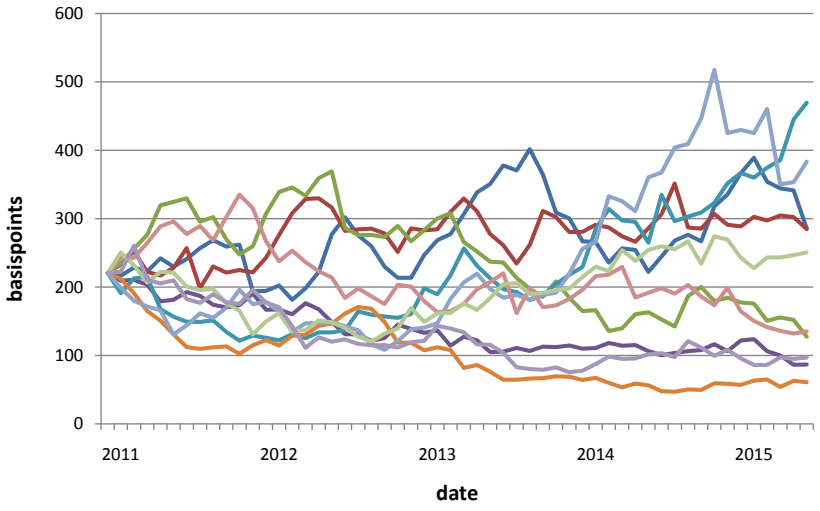


Figure D.1: Graph of 10 different realisations of the retail spread