U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

# Towards a Dynamic Model for Credit Risk

by

Maria do Carmo da Rocha Sousa

Ph.D. Thesis in Business and Management Studies, branch of Finance

Supervised by:
João Manuel Portela da Gama
Elísio Fernando Moreira Brandão

December 2015

# Biographical note

Maria was born in Penafiel, Portugal, in 1980, the youngest of four daughters. She received her BSc in Mathematics, in 2003, and her MSc in Mathematical Engineering, in 2007, both awarded from the Faculty of Sciences of the University of Porto. She started her PhD course in 2010 at the Faculty of Economics of the University of Porto.

In 2002, Maria started to work as a mathematics teacher. In 2004, she began her career in banking, as a risk analyst. By then, she joined the Change the Bank Projects in Millennium bcp, including the "Operational Centralization of OTP/NSF"[1]. Through 2009 and 2010 she participated in the application of the bank's IRB (Internal Rating Based) approach, in Basel II. Until 2012, she also worked with the operations of the BCP group abroad in the role of models analyst, contributing to the IRB application in Poland. In 2013 she was promoted to technical coordinator of a Models and Analytics team in the Credit Risk Department, being responsible for the bank's retail PD models management. Since the end of 2015 she is with the team of Corporate Financing.

In 2006, she received an award of merit by the Banking Services Committee for the contribution of her team "Operational Centralization of OTP/NSF" to the business. In 2013, she won an international competition in Data Mining and Finance[2], as part of the conference BRICS Countries Congress on Computational Intelligence.

During her PhD she presented her work in the BRICS-CCI & CBIC 2013 Countries' Congress on Computational Intelligence, in the 4th International Conference on Information and Finance - ICIF 2014, and in the EFMA Annual Meetings 2015. Part of her research is published in the Journal of Economics, Business and Management and in the Journal of Expert Systems with Applications. She acted as a referee for Evaluation and Program Planning and Neurocomputing Journals.

Maria is married and is mother of two twin boys.

[1] OTP: Outstanding Transaction Posting; NSF: Non-Sufficient Funds.

[2] There were 60 teams registered and 4 finalists in the competition. The other finalists were the University of Technology and Economics of Budapest (Hungary), STATISTICA (Norway) and the Federal University of Rio de Janeiro (Brazil). The competition, open to academy and industry (financial, rating and analytics), focused on the application of credit scoring in the retail banking sector, and its goal was to promote innovative solutions to deal with the temporal degradation of credit rating models.

# Acknowledgements

This work became possible due to the support of several different people and organizations. In particular, I would like to thank to Manuel Gonçalves and to Nelson Teixeira on behalf of the BCP Group, for their financial support. I am grateful to my supervisors, Professors João Gama and Elísio Brandão for their guidance over the last few years. I would like to acknowledge the anonymous referees, Professor Valerio Pesic and Professor Massimiliano Barbi for their constructive comments, discussion and reviews. Thanks to my family, the best in the world! Thanks to Bino and Tina, the most outstanding Grandpa and Nana. A special thank you goes to Jaime for his unlimited protection, patience and care, and to Pedro and Tomás, my little babies, for being the driving motivation in this long journey.

# Abstract

This thesis addresses the problem of credit risk assessment under changing conditions in financial environments. There are a number of serious challenges in this area that have not received sufficient attention so far. This includes adapting the rating systems to population drifts and stress-testing decision systems under extreme adverse conditions.

Research evolves from a baseline methodology to new dynamic modelling frameworks, and is settled in two interconnected research domains: the first assumes that the rating systems' frameworks should adapt to the changing conditions; the second deals with the influence of extreme circumstances in credit default and in the banking business. As part of our contributions, we propose new methodological frameworks for risk assessment, and we present renewed empirical measures for credit risk. Empirical studies use real-world financial databases, of which the most relevant is the Freddie Mac's loan-level dataset, available online since 2013 in the follow-up to the subprime crisis in the United States (U.S.).

In the first research domain we propose an adaptive modelling framework to investigate the two mechanisms of memory, short-term (STM) and long-term memory (LTM), in credit risk assessment. These components are fundamental to learning, but have been overlooked in credit risk modelling frameworks. We argue that different memory windows can be explored concurrently. This is important in the credit risk area, which often undergo shocks. During a shock, limited memory is important; at other times, a larger memory may be preferred.

In the second domain, we have developed a stress-testing methodology under the international rules on capital requirements, the Basel Accords. We present the first study using Freddie Mac database, which describes and implements a risk-adjusted equity model to fund the loans lending. In this context, we analyze the impact of the probability of default (PD) and loss given default (LGD) in the return on lending, when subject to the most extreme adverse circumstances of the past. We contribute with a more realistic understanding of the behavior of the reference static models when exposed to major disturbances in the financial systems.

# Resumo

Esta tese aborda o problema da alteração das condições nos ambientes financeiros para avaliação de risco de crédito. Há vários desafios importantes nesta área que ainda não receberam reflexão suficiente, tais como a adaptação dos sistemas de *rating* na presença de *drifts* e a compreensão do impacto de condições extremas nos sistemas de decisão.

A investigação evolui a partir de uma primeira metodologia de base para novas estruturas de modelagem dinâmicas. O trabalho desenvolveu-se em dois domínios de investigação interligados: o primeiro assume que os quadros dos sistemas de *rating* devem adaptar-se às condições em cada momento; o segundo trata da influência das circunstâncias adversas extremas no incumprimento do crédito e no negócio bancário. Como parte das nossas contribuições, propomos novos quadros metodológicos na avaliação do risco e apresentamos medidas empíricas renovadas para o risco de crédito. Os estudos de caso utilizam bases de dados financeiros reais, das quais a mais relevante é a base de dados da Freddie Mac. Esta tem vindo a ser disponibilizada *online* desde 2013 no seguimento da crise do *subprime* nos Estados Unidos (EU).

No primeiro domínio da investigação, propomos uma *framework* de modelagem adaptativa para compreender os dois mecanismos de memória, de curto e de longo prazo, na avaliação de risco de crédito. Estas componentes são fundamentais para a aprendizagem, mas têm sido negligenciadas nos quadros de modelização atuais. Defendemos que janelas de memória diferentes podem ser exploradas simultaneamente, o que é importante na área do risco de crédito, frequentemente sujeita a choques. Nas crises, a memória curta é importante; em fases estáveis, é desejável uma memória longa.

No segundo domínio da investigação, desenvolvemos uma metodologia de testes de esforço ao abrigo das regras internacionais sobre os requisitos de capital. Apresentamos o primeiro estudo suportado na base de dados da Freddie Mac, que descreve e implementa um modelo de retorno dos empréstimos ajustado pelo risco dos ativos alocados a essas operações. Neste contexto, analisamos o impacto da probabilidade de incumprimento (PD) e da perda dado o incumprimento (LGD) no retorno dos empréstimos, quando sujeitos às circunstâncias adversas mais extremas do passado. Contribuímos assim para uma compreensão mais realista sobre o comportamento dos modelos estáticos de referência quando sujeitos a perturbações nos sistemas financeiros.

# Contents

# List of tables

# List of figures

# List of acronyms

| | |
|---|---|
| AIR | Average interest rate |
| AUC | Area under the ROC curve |
| BCBS | Basel Committee on Banking Supervision |
| BIS | Bank for International Settlements |
| BRICS | Association of five major emerging economies: Brazil, Russia, India, China, and South Africa |
| CRD | Capital Requirements Directive |
| EAD | Exposure at Default |
| EBA | European Banking Authority |
| EL | Expected Loss |
| EU | European Union |
| FED | Central banking system of the United States (also known as the Federal Reserve System or Federal Reserve) |
| FHFA | Federal Housing Finance Agency |
| FHLB | Federal Home Loan Bank |
| FHLMC | Federal Home Loan Mortgage Corporation |
| FIFO | First in First Out |
| FSA | Federal Services Authority |
| FOMC | Federal Open Market Committee |
| GAM | Generalized Additive Models |
| GLM | Generalized Linear Models |
| HAMP | Home Affordable Modification Program |
| HHI | Herfindahl-Hirschman Index |
| IRB | Internal Rating Based |
| IV | Information Value |
| LGD | Loss Given Default |
| LM | Loss Matrix |
| LTM | Long-term memory |
| LTV | Loan to Value |

| | |
|---|---|
| M | Maturity |
| OECD | Organisation for Economic Co-operation and Development |
| PD | Probability of Default |
| PIT | Point-in-time |
| PSI | Population Stability Index |
| RBP | Risk-Based Pricing |
| ROC | Receiver Operating Characteristic |
| STM | Short-term memory |
| TTC | Through-the-cycle |
| UK | United Kingdom |
| US | United States |
| WOE | Weight of Evidence |

# 1. Introduction

More than half a century has passed since credit scoring models were introduced to credit risk assessment and corporate bankruptcy prediction (Harold Bierman and Hausman, 1970, Altman, 1968, Smith, 1964, Myers and Forgy, 1963). Nowadays, in advanced economies, a high proportion of loans are automatically decided using frameworks where the credit score is the central, if not the unique, indicator of the borrowers' credit risk. In the United States (US) the measure of risk FICO score is an industry standard, claimed to be used in 90% of lending decisions, to determine how much capital each individual can borrow and to set the interest rate for each loan. In the OECD countries, probability of default models, or PD models, are the cornerstone for the calculation of banks' regulatory capital in the internal ratings-based (IRB) approach of the Basel revised Framework. For the retail exposures, the determination of the borrower's PD, which in association with the loss given default (LGD) and the exposure at default (EAD), form the key input risk parameters for the minimum capital calculation.

A credit scoring model is an intelligent system. The output is a prediction of a borrower falling into default in the future. Typical credit scoring models are developed from static datasets. Subject to context specifics, and provided that certain requirements of the methods are met, a timeframe for the development is delimited somewhere in the past. Then, by looking at historical examples within such timeframe, the model is designed using a supervised learning approach. The resulting model is then used, possibly for several years, without further adjustments. As a result, these models are quite insensitive to changes in the financial environments, such as population drifts in periods of major financial distress.

Our research has a three-fold objective. First, we attempt to introduce time changing economics and market conditions into credit scoring models. Second, we aim to outline a new and a more realistic modelling framework capable of self-adjusting and learning adaptively from the dynamics that distress or change consumers’ behaviour and their credit worthiness. The third objective is to evaluate the impact of extreme events on credit risk measures, and therefore in banks’ lending outcomes, under the current capital rules.

We study and propose new dynamic modelling frameworks for credit risk assessment. Research is settled in two connected lines: the first assumes that rating systems should adapt to the evolving landscapes; the second deals with the influence of default rates’ fluctuations on banking business.

In our research we start by documenting the underlying problem by focusing on the analysis of an extensive database that is the main driver of the ensuing study on dynamic models, the Freddie Mac’s single loan-level dataset, first published in March 2013. We investigate the dynamics and performance of over 16.7 million of fully amortized, 30-year fixed-rate mortgages in the US, granted between 1999 and the first quarter of 2013. In this investigation, we identify the frailties of the frameworks used in default prediction, to identify the implications to risk-based pricing and rating systems designs. Our analysis shows that, not only scores diminished their ability to predict defaults when the mortgage crisis had come to public’s attention, but also the real default rates by score are quite irregular over time. We therefore conclude that there is a link between scores, lending and default, mostly influenced by the lending practices, although the mapping between these is not adequate. Existing credit scoring models are effective to sort individuals by risk, but they are not suitable to predict real default in each point in time (Sousa et al., 2015a).

In light of the above, the core of our research focuses on investigating and developing an initial approach for dynamic credit scoring modelling, suitable to deal with the temporal degradation of static credit scoring models (Sousa et al., 2013). The model is firstly developed with a classical framework using a static supervised learning and a binary output. Then, using a linear regression between the macroeconomic data and the internal default in the portfolio, the output is adjusted by macroeconomic factors. This methodology jointly deals with the specific risk, which is captured from the data gathered by the time of the application, and the systemic risk, which is modelled with the regression. When applied to a portfolio of customers with credit cards, in a financial institution operating in Brazil, it produced motivating results over a one-year period. Subsequently, we extend this line of investigation by proposing a new dynamic modelling framework for credit risk assessment. The proposed model mimics the principle of films, in that it composes the credit scoring with a sequence of snapshots, rather than a single photograph. Within the new modelling scheme, predictions are made upon the sequence of temporal data, which is suitable for adapting to real default concept drifts, translated by changes in the population, in the economy or in the market. The new model enabled us to depict the two basic mechanisms of memory - short-term (STM) and long-term memory (LTM) - in credit risk assessment. Central to this line of investigation is the idea that models can be improved when acting similarly to human learning and that STM and LTM are the driving components of that process.

The first line of our research argues that the credit risk assessment should evolve with the changes in the conditions underlying the models. However, when rating systems do not have such an ability, as those which prevail in the industry, banks' lending policies and supervisory rules should be oriented by realistic evaluations of the financial business models. With this motivation, we developed a research pipeline, consisting of a stress-testing methodology, which is an important risk management tool used by banks and supervisors.

In the second domain of our research, we present a first application of Freddie Mac's database to stress-testing. Here, we propose a first implementation of a return on risk-adjusted equity model embedded in the contemporary capital regulatory framework, which can easily be replicated to other real-loans portfolios. Our research is pioneer in relating the FICO credit score, which from 1990 is the key driver of lending decision-making in the US, with the return on lending, through the last financial crisis. We found that much of the disturbances on the return on lending occur when the Loss Given Default (LGD) is sufficiently high. Below LGD=25%, empirical simulations show that lending to borrowers with lower scores produces positive returns in the long-run. Therefore, we claim that, for certain values of LGD, credit-cuts by score might have been an overreaction to the crisis. If sufficiently mature, these loans can boost portfolios compositions, because they are less exposed to early payments.

## 1.1. Contributions

The main contribution of this research is a new modelling framework for credit risk assessment that extends beyond the prevailing credit scoring models built upon historical data static settings. The proposed modelling scheme is able to adapt more suitably to changes in the financial environments, like population drifts and shocks in the economy or in the markets.

Our theoretical model is supported by empirical evidence. When the new modelling framework is applied to a set of 16.7 million mortgage loans granted in the U.S. from 1999 through 2013, it becomes apparent that new data consistently improves forecasting accuracy. We claim that existing models are improved when acting similarly to human learning, in which the STM and LTM are key mechanisms of cognition.

While the previous line of investigation provides convincing results, there are some real business problems with adjusting models over time. First, lenders have little incentive to employ such models, because it is expensive and time-consuming to build new scorecards. Models need to be internally tested and validated and then regulators need to approve them. Furthermore, regulators still promote models whose coefficients do not change over time.

In this scenario, stress-testing methods are another important risk management tool used by banks and supervisors to measure the impact of some disorderly events on real systems (like financial crashes), although banks are still using unchanged models. We present a first application of Freddie Mac's database to stress-testing, which can easily be replicated to other real-loans portfolios. Our proposal includes a first implementation of a return on risk-adjusted equity model embedded in the contemporary capital regulatory framework. Under this setting, we analyse the impact of the Probability of Default and the Loss Given Default (LGD) in the return on lending under the most extreme adverse circumstances of the past. We find that under certain values of LGD, lending to lower scored borrowers can produce positive returns in the long run. When sufficient mature, these loans can improve portfolios composition because they are less exposed to early repayments.

## 1.2. Publications and awards

Parts of the second appear in the paper "A new dynamic modeling framework for credit risk assessment", which is published in the Journal of Expert Systems with Applications[1].

The material in chapter 3 is published, on the whole, as a regular paper in the Journal of Economics, Business and Management[2] with the title "Links between Scores, Real Default and Pricing: Evidence from the Freddie Mac's Loan-level Dataset".

Chapter 4 describes a first methodology for dynamic modelling which was used in the model that won[3] a competition in the area of data mining and

[1] ISSN: 0957-4174; DOI:10.1016/j.eswa.2015.09.055. According to the ISI Thomson Reuters Journal Citation Report 2015, the journal impact factor was 2,24 in 2014, being in the first quartile in three research areas: Operations Research & Management Science (12/81), Computer Science, Artificial Intelligence (29/123) and Engineering, Electrical & Electronic (48/249).

[2]JOEBM 2015 Vol.3(12): 1106-1114 ISSN: 2301-3567; DOI:10.7763/JOEBM.2015.V3.343.

[3]http://www.inesctec.pt/ip-en/news-events/news/liaad-inesc-tec-conquista-1o-lugar-em-

finance, promoted by the conference BRICS[4] Countries Congress (BRICS-CCI) on Computational Intelligence, in 2013. It was published in the proceedings of the conference with the title "A two-stage model for dealing with temporal degradation of credit scoring"[5]. A related version is published as a FEP working paper[6] with the title "Introducing time-changing economics into Credit Scoring".

Chapter 5 is a large extension of the work presented in Chapter 4. Our paper "A new dynamic modeling framework for credit risk assessment" presents the methodology, and is published in the Journal of Expert Systems with Applications[7]. Another paper describing an application of this method stands on the Freddie Mac's loans database and was submitted to the Journal of Risk Model Validation with the title "Dynamic credit score modelling with short-term and long-term memories: the case of Freddie Mac's database", and is being revised taking in consideration the reviewers' comments.

The content of chapter 6 was presented in the paper entitled "Stress-testing the return on lending under real extreme adverse circumstances" at the EFMA meetings 2015. Taking into consideration the comments received, the paper was submitted with minor changes to the Journal of Credit Risk, with the title "Stress-testing the return on lending in the aftermath of the crisis under the Basel Capital Rules", where it is under review.

competicao-de-financas-e-data-mining/.

[4]BRICS is the acronym for an association of five major emerging national economies: Brazil, Russia, India, China, and South Africa.

[5] arXiv:1406.7775.

[6] ISSN: 0870-8541; working paper n. 513.

[7] ISSN: 0957-4174; DOI:10.1016/j.eswa.2015.09.055.

## Structure of the dissertation

The remainder of this document is organized in six chapters.

Chapter 2 provides some relevant technical background, which includes the state-of-the-art and base concepts in credit risk modelling.

Chapter 3 outlines the real financial problem motivating our research. Complementing the previous works, we provide a fine-graded time-analysis over scores and analyse in which conditions risk-based pricing and credit decision-making had been implemented. We document the dynamics and performance of over 16.7 million mortgage loans that were at the epicentre of the last global crisis, to find that credit scoring models are effective to sort individuals by risk, but they are unsuccessful in predicting real default to each point in time.

Chapter 4 describes a first approach to dynamic credit scoring modelling, designed to adapt predictions to each point in time. We propose a methodology that jointly deals with the specific risk and the systemic risk, assumed to be time-variant. Firstly, we develop a model with a static supervised learning model and then we adjust the output by a set of public macroeconomic factors.

In chapter 5 we add more depth to the credit score learning. We present an empirical study relying on the Freddie Mac's database where we apply a new dynamic modelling framework for credit risk assessment. In this study we employ two learning configurations of memory, short-term memory (STM) and long-term memory (LTM), to argue that current rating frameworks can be optimized with a dynamic modelling with concurrent STM and LTM.

Chapter 6 describes a first implementation of a return on risk-adjusted equity model under the current capital requirements. The empirical study relying on the Freddie Mac's database assumes the view of an Advanced bank, under the Basel III rules and considering two risk-weighting approaches, as in the U.S. banks and as in the European Union.

Thesis conclusions and paths for future work are discussed in chapter 7.

Some sections are repeated through the chapters of the dissertation to make them self-contained.

# 2. State-of-the-art and base concepts

Parts of this chapter are published in the Journal of Expert Systems with Applications in the paper titled “A new dynamic modeling framework for credit risk assessment”. ISSN: 0957-4174; DOI:10.1016/j.eswa.2015.09.055.

Some sections are repeated through the dissertation to make the chapters self-contained.

## 2.1. Credit scoring - from Fisher’s discriminant to machine learning

The World War II promoted the first expert systems to evaluate a person’s credit worthiness. As credit analysts were called to fight, finance houses and mail-order firms requested them to write down their rules for deciding whom to give loans. Some of these were numerical scoring systems and others were sets of conditions that needed to be satisfied – expert systems. In the early 1950, Bill Fair and Earl Isaac created the first consultancy directed to finance houses, retailers and mail orders firms, making use of statistically derived models in lending decision. Still, until 1970 credit risk assessment relied most exclusively in human judgement. Connected with lending activities, this task was typically performed to support decision-making, following a specific credit application. The labour of the risk evaluator, often a branch manager, would involve the analysis of the likelihood of a customer repaying his debt, based on a number of clues that he could gather from a community leader or an external entity, such as the employer, a credit bureau or another lender. Main aspects that he would check would concern to the customer character and honesty and his ability to create wealth. A person with little or no history of credit, or without proven capacity to gather assets would have little or no chance of having credit. The depth of reasoning behind a decision could largely vary and final decision would much depend on the evaluator's mood and instinct. From customer application to the decision or credit granting, the process was usually slow (Thomas et al., 2002).

The first approach to differentiate between groups took place in Fisher’s original work (1936) for general classification problems of varieties of plants.

The objective was to find the best separation between two groups, searching for the best combination of variables such that the groups were separated the most in the subspace. Durand (1941) brought this methodology to finance for distinguishing between good and bad consumer loans. The first method used in the development of credit scoring systems was discriminant analysis, mainly focused on credit granting in two categories of loans: consumer loans, and commercial loans. For an early review and critique on the use of discriminant analysis in credit scoring, see Eisenbeis (1978).

The boom of credit cards demanded the automation of the credit decision task and the use of better credit scoring systems, which were doable due to the growth of computing power. The value of credit scoring became noticed and it was recognized as a much better predictor than any other judgmental scheme. Logistic regression (Steenackers and Goovaerts, 1989) and linear programming (see Chen et al. (2013) for a review) were introduced in credit scoring, and they turned out to be the most used in financial industry (Anderson, 2007, Crook et al., 2007). The use of artificial intelligence techniques imported from statistical learning theory, such as classification trees (Breiman et al., 1984, Quinlan, 1986) and neural networks (Malhotra and Malhotra, 2002, West, 2000, Desai et al., 1996, Jensen, 1992) have arisen in credit scoring systems. Support Vector Machine (SVM) is another method based in optimization and statistical learning that received increased attention over the last decade in research in finance, either to build credit scoring systems for consumer finance or to predict bankruptcy (Min and Lee, 2005, Li et al., 2006). Genetic algorithms (Chen and Huang, 2003), colony optimization (Martens et al., 2007), and regression and multivariate adaptive regression splines (Lee and Chen, 2005) have also been tried.

The choice of a learning algorithm is a difficult problem and it is often based on which happen to be available, or best known, to the user. The number of learning algorithms is vast. Many frameworks, adaptations to real-life problems, intertwining of base algorithms were, and continue to be, proposed in the literature, ranging from statistical approaches to state-of-the-art machine learning algorithms, parametric to non-parametric procedures. Lessmann et al. (2015) compare several novel classification algorithms to the state-of-the-art in credit scoring and provide an assessment of recent scoring methods that sets a new baseline to which future approaches can be compared. Practice actually suggests that rational behind choice is likely to be more shifted to palatability rather than to accuracy, translated by any performance measure (such as error rate, Receiver Operating Characteristic curve, Gini index, or other).

As an alternative to using a single method, a trend that is still evolving relates to the use of hybrid systems (Lee et al., 2002), and ensemble of classifiers with which the outputs are achieved by a predefined sequence or rule, or a voting scheme. New concepts for adapting to changes (Adams et al., 2010, Pavlidis et al., 2012, Yang, 2007, Jung et al., 2015) and modelling the dynamics (Crook and Bellotti, 2010) in populations start being exploited in credit risk assessment.

## 2.2. Base concepts and techniques

### 2.2.1. Score formulation

A credit scoring model is a simplification of the reality. The output is a prediction of a given entity, actual or potential borrower, entering in default in a given future period. Having decided on the default concept, conventionally a borrower being in arrears for more than 90 days in the following 12 months, those cases matching the criterion are considered bad and the others are good. Other approaches may consider a third status, the indeterminate, between the good and the bad classes, e.g. 15 to 90 days overdue, for which it may be unclear whether the borrower should be assigned to one class or to the other. This status is usually removed from the modelling sample; still the model can be used to score them.

Applied to credit risk assessment, we are essentially considering a supervised learning problem with the aim of predicting the default $y \in$ {good, bad}, given a set of input characteristics $\mathbf{x} \in \mathbf{X}$. The term example, or record, is used to refer to one pair of ($\mathbf{x}$, y). Supervised learning classification methods try to determine a function that best separates the individuals in each of the classes, good and bad, in the space of the problem. A robust model enables an appropriate differentiation between the good and the bad classes. It is achieved by capturing an adequate set of information for predicting the probability of the default concept (i.e. belonging to the bad class), based on previous known default occurrences.

The model building is carried on a set of examples – training set – collected from the past history of credit, for which both $\mathbf{x}$ and y are known. The best separation function can be achieved with a classification method. These methods include, among others, well-known classification algorithms such as decision trees (DT), SVMs, artificial neural networks (ANN), and Generalized Additive Models (GAM). Hands-on software packages are available to the user for example in R, SAS, Matlab, and Model Builder for Predictive Analytics. The accuracy of such functions is typically assessed in separate sets of known examples – test, validation or out-of-sample datasets. The idea is to anticipate the accuracy of that function in future predictions of new examples where **x** is known, but y is not.

The output of these models is a function of the input characteristics **x**, which is most commonly referred as score, $s(\mathbf{x})$. This function has a monotonic decreasing relationship with the probability of entering in default (i.e. reaching the bad status). The notation of such probability is

$$p(B|s(\mathbf{x})) = p(B|s(\mathbf{x})) = p(B|s(\mathbf{x}), \mathbf{x}) = p(B|\mathbf{x}), \forall \mathbf{x} \in \mathbf{X}, \tag{1}$$

where B represents the bad class, or in other words $p(B) = p(y = \text{bad})$. Likewise, G represents the good class and $p(G) = p(y = \text{good})$.

Since $p(G|x) + p(B|x) = 1$, the probability of the complementary class comes as

$$p(G|s(\mathbf{x})) = P(G|\mathbf{x}) = 1 - p(B|\mathbf{x}), \forall \mathbf{x} \in \mathbf{X}. \tag{2}$$

Among researchers and practitioners, a usual form of the score is the log odds score

$$s(\mathbf{x}) = \ln \frac{p(G|\mathbf{x})}{p(B|\mathbf{x})} \text{ and } p(B|\mathbf{x}) + p(G|\mathbf{x}) = \mathbf{1} \tag{3}$$

And so, the score may vary from -∞, when $p(G|\mathbf{x}) = 0$, to +∞, when $p(G|\mathbf{x}) = 1$, i.e $s(\mathbf{x}) \in \mathbb{R}$ . In this case, the probability of the default event can be written in terms of the score as

$$p(B|\mathbf{x}) = \frac{1}{1+e^{s(\mathbf{x})}} \text{ , } \forall \mathbf{x} \in \mathbf{X}. \tag{4}$$

The most conventional way to produce log odds score is based in the logistic regression. However, other classification algorithms can also be used, adjusting the output to the scale of that function. Although a grounded mathematical treatment may be tempting to tackle this problem, it goes beyond the scope of this thesis. The basics of credit scoring and the most common approaches to build a credit scoring model are further detailed in the operational research literature (Thomas et al., 2002, Anderson, 2007). Recent advances in the area deliver methods to build risk-based pricing models and methodologies towards the optimization of the profitability to the lenders (Einav et al., 2013).

In the following sections, we will describe some of the most widely used techniques in credit scoring development, being the Generalized Additive Models (GAM) the core method of our research.

### 2.2.2. Generalized linear models

Linear regression models represent the linear relationship between a continuous response variable and one or more predictor variables (either continuous or categorical) in the form $\mathbf{y} = \mathbf{xw} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the vector of observations of the response variable, $X$ is the matrix determined by the predictors, $\boldsymbol{w}$ is the vector of parameters, $\boldsymbol{\varepsilon}$ is a vector of random disturbances, independent of each other and usually having a normal distribution. These models are appropriated for linear relationships between the response and one or more predictors. The $X$ matrix (and the $\mathbf{w}$ vector) is usually extended with an additional variable, with all data points taking the same constant value on this variable. This allows obtaining linear relationships without the restriction of passing through the origin. The hypothesis of normally distributed regression errors is very restrictive. Generalized linear models (GLM), formulated by Nelder and Wedderburn (1972) are an extension of linear regression models, expanding the use of regression analysis beyond disturbances with normal distribution, which can be used when in the presence of nonlinear relationships.

To understand GLM, first notice that the linear models have the following three characteristics:

- The response has a normal distribution with mean $\mu$.
- A coefficient vector $\mathbf{w}$ defines a linear combination $\mathbf{x}^{\mathrm{T}}\mathbf{w}$ of the predictors $\mathbf{x}$.
- The model equates the two as $\mu = \mathbf{xw}$.

In GLM, these characteristics are generalized as follows:

- The response has a distribution that can be normal, binomial, Poisson, gamma, or inverse Gaussian.
- A coefficient vector $\mathbf{w}$ defines a linear combination $\mathbf{xw}$ of the predictors $\mathbf{x}$.
- A link function $f()$ defines the link between the two as $f(\mu) = \mathbf{x}$.

The most important case in which linearity is not enough is when y and $\mu$ are bounded. The linear model is inadequate in these cases because complicated and unnatural constraints on **w** would be required to make sure that $\mu$ stays within range. Typically, the link function is used to transform the $\mu$ into a scale on which it is unconstrained. The identity link specifies that the expected mean of the response variable is identical to the linear predictor, rather than to a non-linear function of the linear predictor. Although other link functions are possible (e.g. probit and the complementary log-log function), some common canonical link functions for a variety of probability distribution are given below in Table 2.1.

Table 2.1: Typical settings in GLMs.

| **Probability distribution** | **Canonical link Function** | **Meaning $f()$** | **Parameter restriction** |
|---|---|---|---|
| Normal | Identity | $f(\mu) = \mu$ | $\mu\ real$ |
| Binomial | Logit | $f(\mu) = log\left(\frac{\mu}{1-\mu}\right)$ | $\mu \in (0,1)$ |
| Poisson | Log | $f(\mu) = log(\mu)$ | $\mu > 0$ |
| Gamma | Reciprocal | $f(\mu) = 1/\mu$ | $\mu > 0$ |

### 2.2.3. Logistic Regression

Logistic regression is an instance of the generalized linear models. It is similar to a linear regression model but it is suited to models where the dependent variable is dichotomous, that is, the dependent variable can take the value 1 with probability $q$ or the value 0 with probability $1$-$q$. This type of variable is called a Bernoulli (or binary) variable.

The independent or predictor variables in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and the response variables is not linear; instead, the *logit* link function is used: $log\frac{\mu}{1-\mu} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$. Or, stated equivalently, $\mu = \frac{e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}}{1+e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}}$.

Fitting a logistic classifier model implies finding estimates of $\mathbf{w}$ that maximize the likelihood of the model (that is, the probability of the data given the model). It can be shown that this model is optimal when both the class-conditional densities $p(\mathbf{x}|G)$ and $p(\mathbf{x}|B)$ are multi-normal with equal covariance matrices, where $G$ and $B$ represent the two target classes. The hyper-plane of all points $\mathbf{x}$ satisfying the equation $\mathbf{w}^{\mathrm{T}}\mathbf{x} = 0$ forms the decision boundary between the two classes; these are the points for which $p(G|\mathbf{x},\mathbf{w}) = p(B|\mathbf{x},\mathbf{w}) = 0.5$.

### 2.2.4. Generalized additive models

Generalized additive models (GAM) are an extension on GLM, introduced by Hastie and Tibshrirani (1986). GAM is often integrated in commercial analytical tools that are used in scorecard designs (FICO, 2006). GAM is a family of powerful and palatable predictive modelling with a wide applicability range, suitable for business applications like credit scoring. It has been documented as a good solution to fulfill the gap between parametric and non-parametric predictive modelling, providing a good trade-off between interpretability and predictive power (Silva and Cardoso, 2015). GAM generalizes the GLM procedure by replacing the linear predictor with a more general version $s_0 + \sum_{j=1}^{p} s_j(x_j)$, where $s_j(.)$ are smooth functions standardized to verify $E\left(s_j(x_j)\right) = 0$ for $j = 1,..,p$ and $p$ representing the number of predictors. The $s_j(.)$ function can be specified parametrically or not.

## 2.3. Critiques to the existing approaches and new challenges

A few limitations to the existing approaches, idealized in the classical supervised classification paradigm, can be traced in published literature (Crook et al., 1992, Gama et al., 2014, Hand, 2006, Lucas, 2004, Thomas, 2010, Yang, 2007):

- The static models usually fail to adapt when population changes.
- Static and predefined sample settings often lead to an incomplete examination of the dynamics influencing the problem.
- Certain assumptions that are implicit to the methods, often fail in real-world environments. These assumptions regard to:
    - *Representativeness* – the standard credit scoring models rely on supervised classification methods that run on 2-years-old static samples, in order to determine which individuals are likely to default in a future fixed period, 1 year for PD models. Such samples are supposed to be representative of the potential credit consumers of the future, the through-the-door population, and sufficiently diverse to reflect different types of repayment behaviour and to

    allow identifying which characteristics best explain differences between individuals that enter in default from those who not. However, a wide range of research is conducted in small samples that are easily available in public literature.
  - *Stability and non-bias* – the underlying distributions of the training and test sets are the same; classes are perfectly defined, and definitions will not change. Not infrequently there are selective biases over time. Simple examples of this occurrence can be observed when a bank launches a new product or promotes a brand new segment of customers, or when macroeconomics shifts abruptly from an expansion to a recession phase.
  - *Misclassification costs* – these methods assume that the costs of misclassification are accurately known, but in practice they are not.
- The methods that are most widely used in the banking industry, logistic regression and discriminant analysis, are associated with some instability with high-dimensional data and small sample size, intensive variable selection effort and incapability of efficiently handling non-linear features.

Part of the criticism to the existing research is focused in a large number of studies that attempt to establish the relative superiority of classification methods (Hand, 2006). In fact, specific features of the problem often render differences irrelevant or unreal when models are applied in practice, and reported gains resulting from in-lab experiments do not, in many cases, translate into real superiority. Likewise, it is not clear whether existing studies and real-world applications follow basic principles of machine learning. Rather than exhaustively trying to find criteria with which an algorithm outperforms another, by some decimals, research should refocus on the problems to capture their essence (Hand, 2006, Thomas, 2010). On the other hand, regulation urges new approaches for suitably dealing with cyclicality, providing incentives for banks to better manage risk and returns over the long-run.

So far, no comprehensive set of research to this end had much impact into practice. In what concerns to credit risk in retail finance, a great deal of sophistication that is needed regards to the introduction of economic factors and market conditions into current risk-assessment systems (Thomas, 2010, Sousa et al., 2013).

Dominant approaches usually stand on static learning models. However, as the economic conditions evolve in the economic cycle (either deteriorating or improving), also varies the behaviour of an individual, and his ability of repaying his debt, hence, default needs to be regarded as time changing. Also the default evolution echoes trends of the business cycle, and related with this, regulatory movements, and interest rates fluctuations. In good times, banks and borrowers tend to be overoptimistic about the future, whilst in times of recession banks are swamped with defaulted loans, high provisions, and tightened capital buffers. The former lead to more liberal credit policies and lower credit standards, the latter promote sudden credit-cuts. Empirical evidence and theoretical frameworks support a positive, and lagged relationship between rapid credit growth and loan losses (Sousa et al., 2015a).

Traditional systems that are one-shot, fixed memory-based, trained from fixed training sets, and static models are not prepared to process the highly detailed evolving data. And so, they are not able to continuously maintain an output model consistent with the actual state of nature, or to quickly react to changes. These are some of the features of classic approaches that put evidence on that the existing credit scoring systems are limited. As the processes underlying credit risk are not strictly stationary, consumers' behaviour and default can change over time in unpredictable ways. There are several types of evolution inside a population, like population drift that translate into changes in the distributions of the variables or their ability to discriminate between default and non-defaulter individuals, affecting credit scoring performance. There is a new emphasis on running predictive models with the ability of sensing themselves and learning adaptively. Advances on the concepts for knowledge discovery from data streams suggest alternative perspectives to identify, understand and efficiently manage dynamics of behaviour in consumer credit in changing ubiquitous environments. In a world in which events are not preordained and little is certain, what we do in the present affect how events unfold in unexpected ways. The new paradigm of forecasting turns out to be looking at hidden streams in the present signal and understand how they will possibly direct an event into the future.

## 2.4. Dynamic modelling for credit default

### 2.4.1. Concept drift in credit default

Credit default is mostly a consequence of financial distress. A person, or a company, is in financial distress when experiencing individual financial constraints or being exposed to external disturbances. Financial constraints in private individuals may result from abrupt or intrinsic circumstances. In the first case, distress is usually an outcome of sorrowful events like unemployment, pay cuts, divorce, and disease. The second is most commonly related to overexposure, low assets, erratic behaviour, or bad management performance. In this paper we tackle the phenomenon of concept drift in credit default, which we now briefly explain.

In the existing literature, concept drift is generally used to describe changes in the target concept, which are activated by transformations in the hidden context (Widmer and Kubat, 1996, Schlimmer and Granger Jr, 1986) in dynamically changing and non-stationary environments. As a result of these transformations, the target concept can shift suddenly or just cause a change in the underlying data distribution to the model. This means that with time, optimal features may drift significantly from their original configuration or simply lose their ability to explain the target concept. For example, a reduction of the minimum LTV (loan to value) tightens the space of possible values, which is noticed with a change in the distribution, and eventually in the credit default concept. When such drifts happen, the robustness of the model may significantly decrease, and in some situations it may no longer be acceptable. Some authors distinguish real concept drift from virtual drift (Gama et al., 2014, Sun and Li, 2011, Tsymbal, 2004). The former refers to changes in the conditional distribution of the output (i.e., target variable) given the input features, while the distribution of the input may remain unchanged. The later refers to gradual changes in the underlying data distribution with new sample data flowing, whereas the target concept does not change (Sun and Li, 2011). Real concept drift refers to changes in $p(y|\mathbf{x})$, and it happens when the target concept of credit default evolves in time. Such changes can occur either with or without a change in $p(\mathbf{x})$. This type of drift may happen directly as a result of new rules for defining the target classes, good or bad, as those settled by regulators, when new criteria for default are demanded to the banks. Examples of these include the guidelines for the minimum number of days past due or in the materiality threshold for the amount of credit in arrears, issued with the previous Basel II Accord. Another understanding of the real concept drift in credit default is associated with indirect changes in the hidden context. In this case, credit default changes when evolving from one stage of delinquency to another. For example, most of the people with credit until five days past due tend to pay before the following instalment, as most of them are just delayers or forgetters. Yet, the part of debtors in arrears that also fail the next instalment are most likely to be in financial distress, possibly as a result of an abrupt or intrinsic circumstance, and therefore they require more care from the bank. When arrears exceed three instalments, the debtor is most certainly with serious financial constraints, and is likely to fail his credit obligations. More extreme delays commonly translate into hard stages of credit default, which require intensive tracking labour or legal actions. Virtual drifts happen when there are changes in the distribution of the new sample data flowing without affecting the posterior probability of the target classes, $p(y|\mathbf{x})$. With time, virtual drifts may move to real concept drifts. Other interpretations can also be found in literature, for describing an incomplete representation of the data (Widmer and Kubat, 1993), and changes in the data distribution leading to changes in the decision boundary (Tsymbal, 2004). According to some authors, other events can also be seen as virtual drifts, like sampling shift (Salganicoff, 1997), temporary drifts (Lazarescu et al., 2004), and feature change (Salganicoff, 1997). As an example of virtual drift, we might consider the credit decision-making along the recent financial crisis. The lenders had to anticipate if a borrower would enter in default in the future (i.e. being bad). Despite of the macroeconomic factors have worsened, employed people with lower debt to income remained good for the lenders, and so they continued to have access to credit. Although we are mostly interested to track and detect changes in the real target concept, $p(\mathrm{y}|\mathbf{x})$, the methodology introduced in this research attempts to cover both real concept and virtual drifts applied to the default concept drift detection and model rebuilding.

### 2.4.2. Methods for adaptation

Traditional methods for building a credit scoring model consider a static learning setting. In so doing, this task is based in learning in a predefined sample of past examples and then used to predict an actual or a potential borrower, in the future. This is an offline learning procedure, because the whole training dataset must be available when building the model. The model can only be used for predicting, after the training is completed, and then it is not re-trained alongside with its utilization. In other words, when the best separation function is achieved for a set of examples of the past, it is not updated for a while, possibly for years, independently of the changes in the hidden context or in the surrounding environment. New perspectives on model building arise together with the possibility of learning online. The driving idea is to process new incoming data sequentially, so that the model may be continuously updated.

One of the most intuitive ideas for handling concept drift by instance selection is to keep rebuilding the model from a window that moves over the latest batches and use the learnt model for prediction on the immediate future. This idea assumes that the latest instances are the most relevant for prediction and that they contain the information of the current concept (Klinkenberg, 2004).

A framework connected with this idea consists in collecting the new incoming data for sequential batches in predefined time intervals, e.g. year by year, month by month, or every day. The accumulation of these batches generates a panel data flow for dynamic modelling. In Finance, it remains unclear whether it is best having a long memory or forgetting old events. If on the one hand, a long memory is desirable because it allows recalling a wider range of adjust to the present situation. A rapid adaptation to changes is achieved with a short window, because it reflects the current distribution of default more accurately. However, for the contrary reason, the performance of models built upon shorter windows worsens in stable periods. In credit risk assessment modelling, this matter has been indirectly discussed by practitioners and researchers when trying to figure the pros and cons of using a through-the-cycle (TTC) or point-in-time (PIT) schema to calibrate the output of the scorecards to the current phase of the economic cycle. For years, a PIT schema was the only option, because banks did not have sufficient historical data series. Since the implementation of the Basel II Accord worldwide, banks are required to store the data of default for a minimum 7-years period and consider a minimum of 5-years period for calibrating the scorecards.

An original idea of Widmer and Kubat (1996) uses a sliding window of fixed length with a data processing structure first-in-first-out (FIFO). Each window may consist of a single or multiple sequential batches, instead of single instances. At each new time step, the model is updated following two processes. In the first process, the model is rebuilt based on the training dataset of the most recent window. Then, a forgetting process discards the data that move out of the fixed-length window.

Incremental algorithms (Widmer and Kubat, 1996) are a less extreme hybrid approach that allows updating the prediction of models to the new contexts.

They are able to process examples batch-by-batch, or one-by-one, and update the prediction model after each batch, or after each example. Incremental models may rely on random previous examples or in representative selected sets of examples, called incremental algorithms with partial memory (Maloof and Michalski, 2004). The challenge is to select an appropriate window size.

## 2.5. Measuring performance

Predictive modelling tries to find good rules (models) for guessing (predicting) the values of one or more variables in a dataset (target) from the values of other variables in the dataset. Our target is the quality of the individual, which can assume two values: bad or good. More than discriminating between these two possibilities, we will be interested in predicting the probability of defaulting (i.e. being bad). The models to use will then yield a scored dataset as a result of their training.

The construction (training) of a model can be optimized to estimate only the probabilities of each class of the target variable, without incorporating any business objectives for which the predictor will be used. In our case the model would try to predict the probability of (not) default (i.e. being good). Next, we would be left with the decision of selecting a score cut-off, where individuals with risk score greater than or equal to a threshold would be accepted; others, below this cut-off would be rejected. This second stage would need to incorporate the adopted measure of business performance, be it profit, loss, volume of acquisitions, market share, etc. The ROC curves are well-suited for this second operation, enabling both to select the best cut-off for a given model and to compare multiple models, detecting dominant models, points of intersection, etc. ROC curves are able to provide a richer measure of classification performance than scalar measures such as accuracy, error rate or error cost. Because they de-couple classifier performance from class skew and error costs, they have advantages over other evaluation measures such as precision-recall graphs and lift curves (Fawcett, 2006). However, as with any evaluation metric, there are some common misconceptions and pitfalls when using them in practice, so using them wisely requires knowing their characteristics and limitations (Hand and Anagnostopoulos, 2013).

A two-step strategy has the benefit of being flexible in regards to changes in the measure of business performance. To accommodate a change in the loss or profit value, only the cut-off needs to be redefined, while the model is kept unchanged. Moreover, after selecting the best working point (cut-off), it is possible to perform a sensibility analysis, investigating how changes in the performance measure affect the performance of the model. It is important to stress that, often, losses or profits cannot be estimated with great certainty by experts on the company. Therefore, this two-step strategy is easily adapted to future changes. However, caution is in order: if the adopted measure of business performance leads to heavily different losses between the different possible errors, the operating point of the model will be strongly shifted away from the point for which it was trained (equal losses), possibly leading to a substantial degradation in performance.

Another strategy would be to incorporate in the construction of the model the adopted measure of business performance. The training of the model would be focused not in the minimization of the misclassification rate but in the optimization of the profit or loss. In this case, the second stage of optimal cut-off selection is (almost) unnecessary[8]. By integrating the business performance in the model construction we expect to attain an ‘optimal’ classifier, tuned for the business criterion.

### 2.5.1. ROC curves and optimal cut-off selection

When designing a classifier we are essentially trying to minimize two types of errors: the error committed in identifying someone as defaulter, class B, when one is in fact a non-defaulter, class G, individual and the opposite type of error of diagnosing someone as non-defaulter when one is in fact a defaulter. A confusion matrix (Table 2.2) can be used to lay out the different errors:

Table 2.2: Confusion matrix C.

| | **Predicted class** | |
|---|---|---|
| **True Class** | *Defaulter*, $\hat{B}$ | *Non-defaulter*, $\hat{G}$ |
| *Defaulter*, $B$ | $p(B,\hat{B})$ | $p(B,\hat{G})$ |
| *Non-defaulter*, $G$ | $p(G,\hat{B})$ | $p(G,\hat{G})$ |

In this confusion matrix, $p(G,\hat{B})$ represents the probability of that model predicts as defaulter and the real class is non-defaulter; the other probabilities follow accordingly. Note that $p\left(B,\hat{B}\right)+p\left(B,\hat{G}\right)=p(B)$, the a priori default

[8] However, not all models allow the incorporation of the loss or profit matrix in the construction process; others use it in a simplified or approximated mode. Therefore, a second stage of tuning the cut-off may reveal appropriate.

probability in the population. Likewise, $p\left(G,\hat{B}\right)+p\left(G,\hat{G}\right)=p(G)$, the a priori non-default probability in the population and naturally $p(B)+p(G)=1$.

We would like to minimize both $p(B,\hat{G})$ and $p(G,\hat{B})$. At one extreme case if our classifier predicts defaulter for any individual we would have $p\left(B,\hat{G}\right)=0$, but a presumably high $p(G,\hat{B})$ (in fact, equal to $p(G)$); at the other end if the trained classifier predicts always non-defaulter we would have $p(G,B)=0$, but a non-zero $p(B,\hat{G})$ (in fact, equal to $p(B)$).

So, designing a classifier resumes to finding the best trade-off between these two types of errors. And the best trade off depends on the costs associated with each decision. Consider a generic loss matrix, LM (Table 2.3).

Table 2.3: Generic loss matrix LM.

| | **Predicted class** | |
|---|---|---|
| **True Class** | *Defaulter* | *Non-defaulter* |
| *Defaulter* | $l_1$ | $l_2$ |
| *Non-defaulter* | $l_3$ | $l_4$ |

It is easy to see that the expected loss, $E[L]$, for a classifier with the confusion matrix C is:

$$E[L]=l_1\times p\left(B,\hat{B}\right)+l_2 p\left(B,\hat{G}\right)+l_3\times p\left(G,\hat{B}\right)+l_4 p(G,\hat{G}) \qquad (5)$$

Now

$$l_1\times p\left(B,\hat{B}\right)+l_2 p\left(B,\hat{G}\right)$$

$$= p(B)\left[l_1 \times \frac{p\left(B,\hat{B}\right)}{p(B)} + l_2 \times \left(1 - \frac{p(B,\hat{B})}{p(B)}\right)\right]$$

$$= p(B)\left[(l_1 - l_2) \times \frac{p(B,\hat{B})}{P(B)} + l_2\right] \quad (6)$$

where $\frac{p(B,\hat{B})}{p(B)} = p(\hat{B}|B)$ is usually known as *sensitivity* or *true positive rate.*

In the same way

$$l_3 \times p\left(G,\hat{B}\right) + l_4 \times p(G,\hat{G})$$

$$= p(G)\left[(l_3 - l_4) \times \frac{p(G,\hat{B})}{p(G)} + l_4\right]$$

$$= p(G)(l_3 - l_4)\left(1 - \frac{p\left(G,\hat{G}\right)}{p(G)}\right) + p(G) \times l_4 \quad (7)$$

where $\frac{p(G,\hat{G})}{p(G)} = p(\hat{G}|G)$ is usually known as *true negative rate or specificity.*

Then, $E[L]$ can be assumed as

$p(B)(l_1 - l_2) \times \text{sensitivity} + l_2 p(B) + P(G)(l_3 - l_4)(1 - \text{specificity}) + l_4 p(G)$ (8)

Summarizing, the loss of a classifier depends *linearly* only on two parameters, the *sensitivity* and *specificity,* weighted by coefficients derived from the loss matrix and the a priori class probability on the population. This means that we can analyse the performance of a model in a 2D space, the (1- *specificity*, *sensitivity*) space.

However, not all points in this space are possible for a model. Having trained a classifier to output the probability of defaulting, we can vary the cut-off parameter (the probability value at which we start declaring a client as defaulter) and, as we do so, trade *specificity* by *sensitivity*. Fig. 2.1 illustrates an example of an ROC chart for two different classifiers. The ROC of a classifier shows this trade-off, plotting the achievable (1-specificity, sensitivity) values for a range of cut-offs.

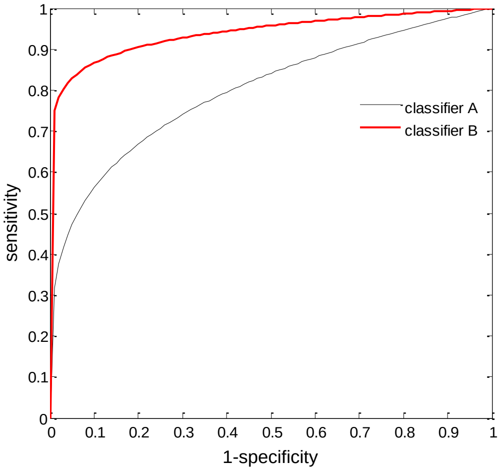


Fig. 2.1: ROC chart for two different classifiers.

Each point on the curve represents a cut-off probability. Points closer to the upper-right corner correspond to low cut-off probabilities. Points in the lower left correspond to higher cut-off probabilities. The extreme points (1,1) and (0,0) represent no-data rules where all cases are classified into class Defaulter or class non-defaulter, respectively. Now, as shown above, the expected loss can be represented as a linear function of sensitivity and 1-specificity:

$$E[L] = a \times (1 - \text{specificity}) + b \times \text{sensitivity} + c. \tag{9}$$

To minimize the loss we just have to walk in the direction opposite to the gradient of $E[L]$. It is not difficult to see that the represented classifier B is dominant over classifier A, in the sense that, for any (reasonable) loss matrix considered, there is always an operation point of classifier B that is better than the best operating point of classifier A.

A different situation is depicted in Fig. 2.2. The ROC curve of classifier A intersects the ROC curve of the classifier C. Now, the best classifier depends on the loss matrix. The isocost line (a isocost line is a line of constant cost, perpendicular to the gradient of the loss function) shows the least loss line for a loss matrix M2 where the costs of missing a positive case severely outweighs the cost of raising a false alarm; in this case classifier C provides the best operating point. Conversely, the dotted isocost line shows the least loss line for a loss matrix M1 where the costs of missing a negative case severely outweighs the cost of raising a positive alarm. Now is classifier A that provides the best operating point as illustrated in Fig. 2.2.

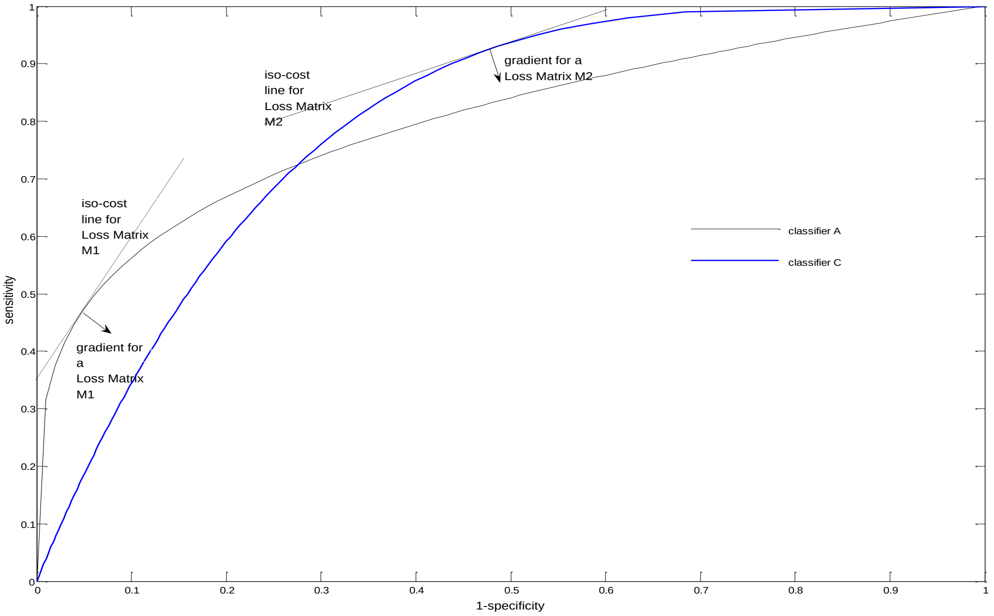


Fig. 2.2: Non-dominant ROCs.

### 2.5.2. Comparing models performance

Throughout this research the experimental comparison of different models (or methods) is based in their ability to discriminate between the two target classes. The discriminatory power is measured with the Gini coefficient, a typical evaluation criteria among researchers and in the industry (Rêzác and Rêzác, 2011), which can be calculated as $2 \times AUC - 1$, where AUC is the area under the ROC curve. The Gini coefficient refers to the global quality of the credit scoring model, and may range between -1 and 1. The perfect scoring model fully distinguishes the two target classes, good and bad, and has a Gini index equal to 1. A model with a random output has a Gini coefficient equal to zero. If the coefficient is negative, then scores have a reverse meaning. The extreme case -1 would mean that all examples of the good class are being predicted as bad, and vice-versa. In this case, the perfect model can be achieved just by switching the prediction. Loans' records with unknown score are not included in the calculations of this indicator.

# 3. Credit scoring models degradation

***Abstract* -** Evidence from the Freddie Mac's single loan-level dataset, first published in March 2013, shows that existing scores are effective to order individuals by risk, but they are not prepared to predict real default in each point in time.

We investigate the dynamics and performance of over 16.7 million of fully amortized 30-year fixed-rate mortgages in the U.S., originated between 1999 and the first quarter of 2013. We identify the frailties of the frameworks used in default prediction, to draw implications to risk-based pricing designs. Analysis shows that not only scores diminished their ability to predict default when the mortgage crisis has come to public's attention, but also that real default rates by score are irregular over time. It is also apparent that, since 2009, lenders are firmly declining the subprime loans, and as a result the 1-year cumulative default by vintage has declined. There is a link between scores, lending and default, mostly influenced by the lending practices. There is a link between scores, default and pricing, but the mapping between them is far from being adequate.

Parts of this chapter are published in the Journal of Economics, Business and Management with the title "Links between Scores, Real Default and Pricing: Evidence from the Freddie Mac's Loan-level Dataset". JOEBM 2015 Vol.3(12): 1106-1114 ISSN: 2301-3567; DOI:10.7763/JOEBM.2015.V3.343.

Some sections are repeated through the dissertation to make the chapters self-contained.

## 3.1. Introduction

The subprime mortgage lending crisis in the U.S. came to public's attention when home foreclosures begun to rise in 2006 and moved out of control in 2007. A large decline in home prices prompted a devaluation housing-related securities and an unprecedented rise in mortgage delinquencies. This brought into light the disproportionate risk assumed in mortgage lending in the last decade, along the bursting of the U.S. housing bubble, between 2001 and 2005. This crisis echoed severely in the financial arena and in real economies worldwide. The collapse of several major financial institutions in 2008, promoted the distrust inside the financial systems. As a consequence, banks' liquidity plummeted with a significant disruption of the financing of businesses and consumers. Thus far, the U.S. and the European Communities are still recovering from a severe recession. This spawned intensive debates towards causes and possible remedies, in view of achieving transparency and global financial stability.

Since 21 March 2013, Freddie Mac[1] is making available loan-level credit performance data on a portion of fully amortized 30-year fixed-rate mortgages that the company purchased or guaranteed since 1999. This had never been done before by a loan level agency. The data is provided in a "living" dataset (Freddie Mac, June 2013a). By June 2014, the dataset covered over 16.7 million of fully amortized, 30-year fixed-rate mortgages in the U.S., originated between 1999 and the first quarter of 2013. These loans represent a total amount granted of over 3,020 US B$. Disseminating these

[1] Freddie Mac is the Federal Home Loan Mortgage Corporation (FHLMC), a public government-sponsored enterprise (GSE) in the U.S. It was created in 1970 to expand the secondary market for mortgages in the U.S.

data follows the direction of the regulator, the Federal Housing Finance Agency (FHFA), as a part of a larger effort to increase transparency and promote risk sharing. The primary goal of turning this data available is to help investors build more accurate credit performance models in support of the risk sharing initiatives highlighted by the FHFA in the 2013 conservatorship scorecard (Federal Housing Finance Agency, 2013). The availability of such a large real world financial dataset also creates an unprecedented opportunity for researchers and practitioners, as it allows a more profound investigation on the roots of the global crisis. The aggregated data summary statistics are updated by Freddie Mac (June 2014b).

Anderson, Scott, and Janet Jozwik (2014) proposed a framework for developing a credit model based on this dataset. For a 180-days delinquent target event, the authors conclude that much of the variation in credit performance across loans and over different stages of the economic cycle is explained by loan-level variables. Unsurprisingly, by adding factors to capture broader macroeconomic effects and the quality of underwriting, they significantly improve the model. Goodman, Landy, Ashworth and Yang (2014) present an exploratory paper providing a first look through the data, to find potential implications for guarantee pricing. The authors show the vintage composition as a percentage of the initial balance in a cross-analysis of the original borrowers’ FICO score by the original loan to value (LTV). They follow the cumulative default in three groups in the score ranges 300 to 700, 700 to 750 and 750 to 850 crossed by the original LTV in selected buckets. They conclude that default rates are dramatically higher on higher LTV/lower scores, and so, investors should look not only at the average LTV and FICO scores, but also at the FICO/LTV loans’ distribution. The authors conjecture that pricing these pools by looking at averages are likely to lead to under-priced default risk, but they do not present evidence.

Discussion is being pushed towards risk-based pricing. Previous studies suggest that risk-based pricing models will rely mostly in credit scores. This research extends the existing published work by proving meaningful insights on the link between credit score, lending practices, real default and pricing. In this chapter we address the question: is there is a link between scores, real default and pricing?

Our research confirms that there is a link between scores, default and pricing, but the mapping between them is far from being adequate. New evidence from the Freddie Mac's single loan level data shows that although existing scores consistently order portfolios' risk, real default rates by score are quite irregular over time. This means that existing scores are effective to order borrowers' risk, but they are not prepared to adapt predictions to real default in each point in time.

This chapter follows in section 3.2 with the formalization of the problem and a description of the research background. In section 3.3, we present an overview of credit scoring models. First, we review the current role of credit scoring in the advanced economies, and then we present credit scoring formulation, for an in-depth understanding. The section ends with a brief explanation of current capabilities and potential frailties of credit scoring models when they are used at the basis of credit risk underwriting and risk-based pricing. Experimental design is explained in section 3.4, and results are provided in section 3.5. Selected outcomes are presented in order to illustrate dynamics over time, and focusing the dimensions in analysis: score buckets, lending practices, observed defaults and pricing. Conclusions are drawn in section 3.6.

## 3.2. Problem and research background

### 3.2.1. The problem

Lenders determine if the risk of lending to a borrower is acceptable under certain parameters of credit risk, borrower's credit capacity and collateral evaluation. Nowadays, in retail lending, the risk of a great proportion of the loan applications is automatically evaluated. In this setting, credit score is the central, if not unique, indicator of the borrowers' credit risk, either when the credit decision assessment is fully automatic or when it is an input for human decision. An individual without a credit score or with a low score (meaning high risk) is unlikely to have credit, whilst an application of a person with a high score has good chances to be accepted.  An analysis on the causes and effects of the mortgage meltdown (Bianco, 2008) states that in 2007, 40% of all subprime loans have been generated by automatic underwritings in the U.S. This had been associated to lax controls in the underwriting processes. The author argues that the automated processes meant fasters decision, but less documentation scrutiny. The acceptance standards have also moved rapidly towards credit score's over-dependence. Hence, the performance of credit loans strongly relies in the credit scoring models accuracy, both in the short-term as in the long-run predictions, which is too hard to achieve. In 2007, the delinquency rate rose sharply, both in borrowers in the lower scores as in the highest scores bands, showing that the actual risk of these borrowers have been underestimated.

Enhancing loans risk-based pricing models in the track of the previous studies (Anderson and Jozwik, 2014, Goodman et al., 2014) will much depend on the knowledge and ability to improve the existing credit scoring robustness. This entails a deeper understanding of their actual strengths and insufficiencies.

### 3.2.2. Research background

To our knowledge, research in credit risk assessment often lacks from validation in representative real world environments, and most of the experimental designs use datasets that are not representative of each phase in the economic cycles. Hence, a significant number of empirical studies have no generalization ability. In particular, trying to screen credit losses and predicting credit default of future credit operations may become critical if there is neither sufficient knowledge of the past neither of the future circumstances. In this setting, theoretical contributions have limited impact in real world decisions.

The unavailability of representative datasets has shortened the space to turn evident in which conditions the existing credit scoring models may be ineffective, like biased credit policies, drifting population and recessions. The single-family mortgage loan level dataset creates an unprecedented opportunity for researchers and practitioners, as long as it enables the simulation of theoretical frameworks in real-world stressed environments.

## 3.3. Credit scoring fundamentals and current use

### 3.3.1. Score - a standard risk assessment in the advanced economies

Financial industry turned over-dependent of credit scoring over the last few decades. The origin of these models traces back to the World War II, which promoted the first expert systems to evaluate a person's credit worthiness. As credit analysts were called to fight, finance houses and mail-order firms requested them to write down their rules for deciding whom to give loans. Some of these were numerical scoring systems and others were sets of conditions that needed to be satisfied – expert systems. In the early 1950s, Bill Fair and Earl Isaac created the first consultancy directed to finance houses, retailers and mail-orders firms, making use of statistically derived models in lending decision. Until 1970 credit risk assessment relied most exclusively in human judgment. Connected with the lending activities, this task was typically performed to support decision-making, following a specific credit application. The labour of the person responsible for the evaluation, often a branch manager, would involve the analysis of the likelihood of a customer repaying his debt, based on a number of clues that the manager could gather on site from a community leader or an external entity, such as the employer, a credit bureau or even another lender. Main aspects that he would check would concern to the customer character and honesty and his ability to create wealth. The depth of reasoning behind a decision could largely vary and the final decision would likely depend on the evaluator's mood and instinct. From customer application to the decision or credit granting, the process was usually slow (Thomas et al., 2002). Nowadays, scoring models are used in credit approval, risk management, internal capital allocation and in corporate governance functions of banks using the IRB approach. In the U.S., since its introduction 20 years ago, FICO score is calculated from the information available in the individuals' credit bureau reports, and has become an industry standard. It is claimed to be used in 90%

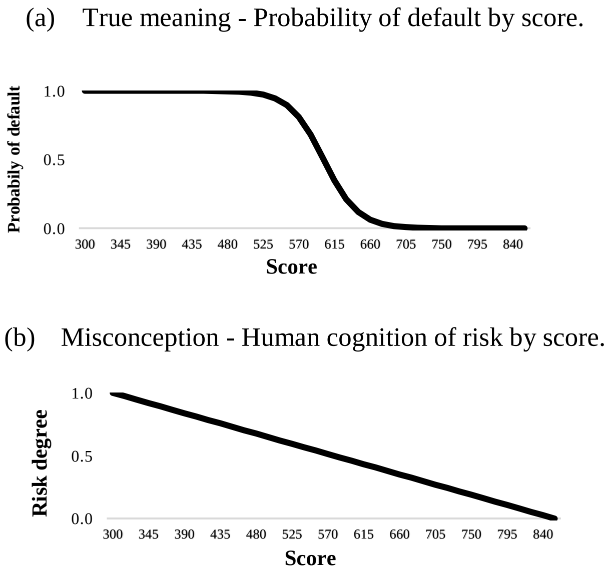
of lending decisions, to determine how much money each individual can borrow, and how much interest he will pay. In the OECD countries, banks that have adopted the Internal Rating Based Approach (IRB) in Basel II, internally developed credit scoring models play an essential role in the calculation of the minimum regulatory capital. In the European Market, there are 89 banks using the IRB. In the U.S., the largest banks also adopted the Basel II Accord, introduced via Capital Requirements Directive. In line with this evolution, financial industry moves toward a more intensive use of credit scores at the basis of risk-based pricing models.

### 3.3.2. Scoring models – strengths and frailties

The huge success of credit scoring models in the advanced economies is partly explained by their appealing representation is a linear scale. Credit scoring models have also proved to a powerful measure to order a population of individuals according to their credit risk. The best scoring model is the one that differentiates the most the two target classes, good and bad, commonly referred as discriminatory power.

### 3.3.2.1. Human misconception

Although the true meaning of scores is a probability of default (PD) with a non-linear shape, as in Fig. 3.1(a), human cognition retains the linear representation, Fig. 3.1(b), rather than the actual non-linear shape. Through our experience in developing credit scoring models it became apparent that many of the risk managers were basing their credit risk assessments on the linear representation. Doing so is suitable for ordering risks, but it is insufficient to calculate losses or pricing credit risks.


(a) True meaning - Probability of default by score.

(b) Misconception - Human cognition of risk by score.

Fig. 3.1. Actual meaning of scores and human misconception of the risk.

Authors' analysis, illustrative scorecard: the score 660 has a good/bad odd of 15 good individuals to 1 bad, and for each additional 15 score points the good/bad odds double.

### 3.3.2.2. Economic cycle and scores misalignment

Traditional systems that are the basis of credit scoring models are one-shot, fixed memory-based, trained from fixed training sets. Since static models are not prepared to process the highly detailed evolving data, they are not able to continuously maintain an output (PD) consistent with the current state, or to quickly react to changes (Gama, 2010). When there are significant changes in the conditions, scores' PD (Fig. 3.2, dashed line) may become misaligned[2] with the observed default (Fig. 3.2, solid line).

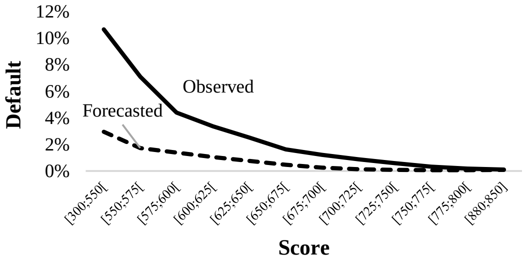


Fig. 3.2. Illustration of scores misalignment.

Authors' analysis based on the Freddie Mac's database. Forecasted PD's by score same as the default rate of the mortgages originated in 1999, and observed PD's the real default rates in 2007. Performance measured in the 1-year after the loans were originated.

As the processes underlying credit risk are not strictly stationary, consumers' behaviour and default can change over time in unpredictable ways. There are several types of evolution inside a population, like population drifts, that translate into changes in the distributions of the variables, affecting the performance of the models.

When the economic conditions evolve in the economic cycle, deteriorating or improving, the behaviour of the individuals, and their ability of repaying their debts also vary. In addition, default evolution echoes trends of the business

[2] The terms alignment, adjustment and calibration are commonly used with the same meaning.

cycle, and related with this, regulatory movements, and interest rates fluctuations. In good times, banks and borrowers tend to be overoptimistic about the future, whilst in times of recession banks are swamped with defaulted loans, high provisions, and capital buffers turn highly conservative. The former leads to more liberal credit policies and lower credit standards, the later promotes sudden credit-cuts. Empirical evidence and theoretical frameworks support a positive, and lagged relationship between rapid credit growth and loan losses (Sousa et al., 2015a).

In order to adapt models' output to changes over time, institutions should calibrate their scoring models according to the most recent information. Models' adjustments, or calibration, commonly consider selected macroeconomic public indicators and should be periodically revised. However, this may take too long to occur. The European Banking Authority reports that there is not a common practice among Regulators towards models calibration. Many countries do not define any specific rules and, when they do, they are usually not public. When they define some rules, they are rarely convergent; and different countries favour different calibration choices (EBA, 2013b). Should there be significant changes in between scheduled modelling developments or adjustments, it is not certain that banks will anticipate any of these tasks, as it can largely depend on judgmental reasoning or over-layered decision frameworks.

In the following, we will analyse the Freddie Mac's database to find evidence on the previous limitations. Complementing the previous works, we provide a fine-graded time-analysis over scores and analyse in which conditions risk-based pricing has been implemented. Furthermore, we study the extent of misalignment of PD's with the real default by score over time.

## 3.4. Experimental design

The research summarized here was conducted in the Freddie Mac's single family mortgage loan-level dataset, first published in March 2013. We follow the performance of 16.737 million of fully amortized 30-year fixed-rate mortgages loans in the U.S., originated between January 1, 1999 and March 31, 2013. The loans performance[3] is outlined in a monthly basis and, at the time of this research, data for performing loans and those that were up to 180 days delinquent were available through September 30, 2013.

The dataset is a "living" dataset updated over time, typically at the end of each quarter, and may be subjected to periodical corrections by Freddie Mac. The release changes  are recorded (Freddie Mac, June 2014a). A general user guide describing the file layout and data dictionary is also provided (Freddie Mac, June 2013b).

Freddie Mac's information regarding the key loan attributes and performance metrics can be linked to our research in the aggregated summary statistics.

[3] Loan performance information includes the monthly loan balance, delinquency status and information regarding termination events: Voluntary prepayments in full; 180 days delinquency ("D180"); Repurchases prior to D180; Third-party sales prior to D180; Short sales prior to D180; Deeds-in-lieu of foreclosure prior to D180; Real estate owned (REO) acquisition prior to D180. Also includes voluntary prepayments and loans that were short sales, deeds-in-lieu of foreclosure, third party sales, and REOs.

### 3.4.1. Methodology

We attempt to describe the most significant events in the period. We are both interested in determining the main contrasts by score, and to illustrate the dynamics over time. First, we illustrate the volumes and compare the original interest rate with the annual average FIX 30[4]. In so doing, our aim is to use a representation of the credit risk spread evaluation over time and understand the extent of under-priced loans that has been referred as one cause of the crisis. Then, we determine the evolution of default over time and the performance of the scores at the basis of the credit risk assessment. For assigning the default event, we used the information of the loan delinquency status in each reporting period. In this analysis we consider that a borrower entered in default if he was ever 90 or more days delinquent, the typical definition used under the Basel II. Default is assigned to the first occurrence of this event. We use vintage analysis and hence, we consider cumulative default along time.

### 3.4.2. Data aggregation

Although we have found missing values in the score, we intentionally kept these cases in the analysis to entirely represent the extent of information (or absence) at the basis of the original risk assessment. Data of the original datasets were aggregated by the origination year.

[4] FIX 30 is the interest rate of a 30-year fixed-rate mortgage.

Scores may vary in the range 301-850, or be unknown. Situations where the score is unknown are described by Freddie Mac (June 2013b)[5]. To compare evolutions over time, we divided the range into equidistant intervals of 25, except for the lower and upper bounds. To have dimension, these bounds were aggregated in the buckets [300, 550[ and [800, 850[, respectively.

### 3.4.3. Scores - performance, concentration and stability measures

The discriminatory power of the model was measured based on the Gini coefficient, equivalent to consider the area under the ROC curve (AUC), which is a typical evaluation criteria among researchers and in the industry (Rêzác and Rêzác, 2011). This coefficient refers to the global quality of the credit scoring model, and may range between -1 and 1. The perfect scoring model fully distinguishes the two target classes, good and bad, and has a Gini index equal to 1. A model with a random output has a Gini coefficient equal to zero. If the coefficient is negative, then the scores have a reverse meaning. The extreme case -1 would mean that all examples of the good class are being predicted as bad, and vice-versa. In this case, the perfect model can be achieved just by switching the prediction. Loans' records with unknown score were not included in the calculations of this indicator.

The concentration of loans by score buckets was measured with the Herfindahl-Hirschman Index (HHI), which is defined as $\sum_{i=1}^{n} {f_i}^2$, where *n* is the number of score buckets and $f_i$ is the number of customers in that bucket

[5] A possible reason is when the seller requires a reduced level of verification.

relative to the total portfolio. By definition, the index varies between 0 and 100%. An index of 100% means that borrowers are concentrated in a single bucket. In this work we will consider that values below 20% are commonly acceptable. Values above it suggests highly concentrated scores.

The stability index was measured comparing the distribution of the population in each year with the distribution of the population in the first year in the period, 1999. For the year $t$, $t = 2000, .., 2013$ the stability index is calculated as:

$$\sum_{i=1}^{n} \left(f_{i,1999} - f_{i,t}\right) * ln\left(f_{i,1999}/f_{i,t}\right) \tag{1}$$

where *n* is the number of score buckets, and $f_{i,t}$ is the number of borrowers in that bucket relative to the number of borrowers in the total portfolio in the year *t*.

## 3.5. Results

There are a number of theories regarding the origins of the mortgage crisis. In this chapter we are concerned to extend knowledge on the potential misalignment of the risk indicators that were at the basis of credit approval before crisis. Then, we assess the extent of undervaluation of the credit risk, and hence, credit risk mispricing. As credit risk assessment is anchored in the borrowers' score at the origin of the loan, the analysis is focused in the dimensions score and time. Results are mostly motivated to present evidence on the following:

- Risk taking and pricing;
- Changes in the lending practices after the crisis;
- Risk assessment over time - default rates, default misalignment and scores performance.

### 3.5.1. Risk taking and pricing

The evolution of new loans over the analysed period illustrates the U.S. housing bubble between 2001 and 2005. The highest peak occur between 2001 and 2003, where the number of new loans continuously rose from nearly 800 thousand loans in 2000 to 1,930 thousands in 2003 (Table 3.1, first row). This massive increase of the number of loans was one of the sources of the raise in the real state property values, which reached a peak in 2005.

Our analysis confirms that scores are intensively used to differentiate the mortgages interest rates. As illustrated in Table 3.2, there is a decreasing trend of the average interest rate from the lower to the higher score buckets. It can be said that a risk-based pricing based in scores is being applied. Until 2009, borrowers with the highest scores were borrowing below the average FIX 30. In 2007, the default of the borrowers with scores 650 or higher has almost tripled in relation to the previous years. Adjustments to the risk premium were made by 2009, as it can be seen in Fig. 3.3. From this point onwards, rates are higher than FIX 30, suggesting that pricing policy had been revised. Loans have been priced below the FIX 30 in 2001, both in the aggregate level and in each score bucket (Table 3.1, rows 7, 8 and 9). Loans' average rate was maintained through 2001 and 2002, in the aggregate level and in score buckets' level (Table 3.2). This effect may be linked to the crash of the dot-com bubble in 2000 which has been associated to the beginning of the decline in real long-term interest rates. In reaction to the crash of the dot-com bubble in 2000 and to the recession that began in 2001, the Federal Reserve Board has cut short-term interest rates from 6.5% to 1% (Bianco, 2008). Mortgage interest rates continued to decline until 2005 (Table 3.2). As the mortgage rates are typically set relation to 10-year Treasury bond yields, this was an outcome of very low Fed funds' rates in the period. Lenders were self-reliant that they were taking little risk because the value of the collateral was rising too fast, but they missed to understand that it would come to an end. Loans underwritten between 2001 and 2005 account for 42% of the amount originated in the period (Table 3.4). There is a theory (Hull, 2009) referring that in this period, lenders had begun to take more risk in subprime[6] mortgages. Our analysis provides a divergent finding, because neither the number of loans nor the amount granted has increased during the mortgage bubble (Table 3.3 and Table 3.4, see rows for scores' buckets bellow score 625).

[6] A rule of thumb for the subprime mortgage is a loan of a borrower with a score inferior to 620. Some lenders also consider a subprime mortgage if the borrower has a score up to 680 and the down payment is less than 5% of the loan.

Table 3.1: Freddie Mac database – loans main indicators, 1999-2013(Q1).

| Indicator | Origination year | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 (Q1) |
| Total loans (thousands) | 1.095 | 787 | 1.757 | 1.685 | 1.930 | 1.131 | 1.324 | 1.083 | 1.069 | 986 | 1.513 | 788 | 556 | 787 | 247 |
| Total original amount (billion US $) | 138 | 104 | 260 | 262 | 311 | 188 | 240 | 202 | 202 | 210 | 345 | 177 | 131 | 192 | 58 |
| Avg original loan amount (‘000 US $) | 126 | 132 | 148 | 156 | 161 | 167 | 181 | 187 | 189 | 213 | 228 | 224 | 236 | 244 | 237 |
| Scores concentration index[6] | 13% | 13% | 13% | 13% | 14% | 13% | 13% | 12% | 12% | 14% | 20% | 19% | 20% | 21% | 20% |
| Scores stability index[7] | n.a. | 0,02 | 0,01 | 0,00 | 0,03 | 0,02 | 0,03 | 0,00 | 0,00 | 0,10 | 0,28 | 0,00 | 0,00 | 0,01 | 0,00 |
| Average interest rate (AIR) (%)[8] | 7,31 | 8,18 | 6,58 | 6,58 | 5,78 | 5,86 | 5,88 | 6,44 | 6,41 | 6,10 | 5,02 | 4,81 | 4,59 | 3,81 | 3,64 |
| AIR-FIX 30 (%) | -0,13 | 0,13 | -0,39 | 0,04 | -0,05 | 0,02 | 0,01 | 0,03 | 0,07 | 0,07 | -0,02 | 0,12 | 0,14 | 0,15 | … |
| AIR at a low score - FIX 30 (%)[9] | 0,02 | 0,33 | -0,14 | 0,29 | 0,12 | 0,16 | 0,14 | 0,17 | 0,25 | 0,49 | 0,44 | 0,56 | 0,51 | 0,46 | … |
| AIR at the highest scores - FIX 30 (%)[9] | -0,17 | 0,04 | -0,50 | -0,07 | -0,09 | -0,02 | -0,04 | -0,02 | -0,02 | -0,04 | -0,07 | 0,05 | 0,07 | 0,12 | … |

…: not available.

[6] We used Herfindahl-Hirschman Index (HHI), for which values below 20% are commonly considered acceptable.

[7] We used population stability index, for which values below 0,25 are commonly considered normal.

[8] Calculated as the weighted average of rates by score buckets.

[9] For low scores we considered the scores in the range [600; 625[; for the highest scores we considered the scores in the range [800; 850[.

Table 3.2: Freddie Mac database – original interest rates, 1999-2013(Q1). Unit %.

| Score bucket | Credit risk assessment | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 (Q1) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unknown | Unpublished | 7,49 | 8,28 | 6,74 | 6,74 | 5,91 | 5,98 | 5,92 | 6,30 | 6,28 | 6,40 | 5,31 | 5,10 | 5,14 | 3,97 | 3,91 |
| [300;550[ | Highest risk | 7,50 | 8,29 | 6,88 | 6,88 | 5,97 | 6,04 | 6,09 | 6,75 | 6,92 | 6,77 | 5,48 | 5,46 | … | 3,71 | 3,62 |
| [550;575[ | ↑ | 7,50 | 8,35 | 6,93 | 6,94 | 6,02 | 6,10 | 6,14 | 6,72 | 6,81 | 6,75 | 5,62 | 5,44 | 5,38 | 4,00 | … |
| [575;600[ | | 7,53 | 8,53 | 7,00 | 7,00 | 6,07 | 6,10 | 6,11 | 6,65 | 6,67 | 6,60 | 5,56 | 5,29 | 4,99 | 3,95 | … |
| [600;625[ | | 7,46 | 8,38 | 6,83 | 6,83 | 5,95 | 6,00 | 6,01 | 6,58 | 6,59 | 6,52 | 5,48 | 5,25 | 4,96 | 4,12 | 3,92 |
| [625;650[ | | 7,40 | 8,29 | 6,73 | 6,73 | 5,89 | 5,95 | 5,97 | 6,53 | 6,53 | 6,48 | 5,44 | 5,22 | 4,97 | 4,12 | 3,95 |
| [650;675[ | | 7,36 | 8,23 | 6,65 | 6,65 | 5,84 | 5,90 | 5,93 | 6,50 | 6,48 | 6,36 | 5,34 | 5,14 | 4,90 | 4,07 | 3,90 |
| [675;700[ | | 7,33 | 8,20 | 6,59 | 6,60 | 5,81 | 5,88 | 5,90 | 6,46 | 6,44 | 6,20 | 5,22 | 5,01 | 4,80 | 3,96 | 3,82 |
| [700;725[ | | 7,30 | 8,16 | 6,56 | 6,56 | 5,78 | 5,86 | 5,88 | 6,44 | 6,41 | 6,13 | 5,11 | 4,89 | 4,69 | 3,87 | 3,71 |
| [725;750[ | | 7,27 | 8,13 | 6,53 | 6,53 | 5,75 | 5,84 | 5,86 | 6,43 | 6,38 | 6,06 | 5,04 | 4,82 | 4,62 | 3,81 | 3,65 |
| [750;775[ | | 7,25 | 8,10 | 6,50 | 6,50 | 5,73 | 5,81 | 5,83 | 6,40 | 6,35 | 6,02 | 5,00 | 4,78 | 4,57 | 3,79 | 3,62 |
| [775;800[ | | 7,26 | 8,08 | 6,47 | 6,47 | 5,72 | 5,81 | 5,81 | 6,37 | 6,31 | 5,98 | 4,97 | 4,76 | 4,54 | 3,77 | 3,60 |
| [800;850] | Lowest risk | 7,27 | 8,09 | 6,47 | 6,47 | 5,74 | 5,82 | 5,83 | 6,39 | 6,32 | 5,99 | 4,97 | 4,74 | 4,52 | 3,78 | 3,60 |
| Weighed average rate | | 7,31 | 8,18 | 6,58 | 6,58 | 5,78 | 5,86 | 5,88 | 6,44 | 6,41 | 6,10 | 5,02 | 4,81 | 4,59 | 3,81 | 3,64 |

… : not applicable.

### 3.5.2. Changes in lending after the crisis

Major drifts have occurred in lending practices in reaction to the crisis, which was noticed both in the acceptance scores thresholds (Table 3.3) as well as in the underlying credit risk spreads (Fig. 3.3). From 2009 onwards, interest rates were increased in all scores, when compared to the average FIX 30 in the year.

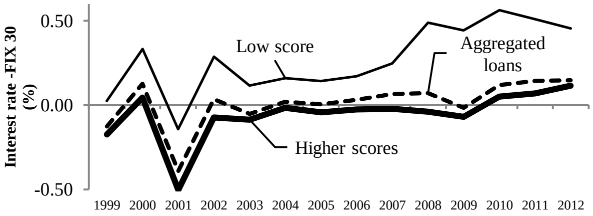
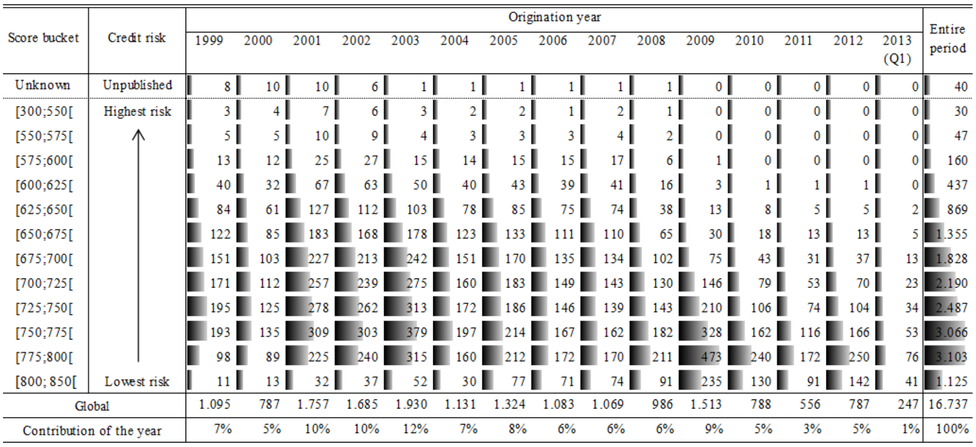


Fig. 3.3. Freddie Mac database - gap between the interest rate and the FIX rate.

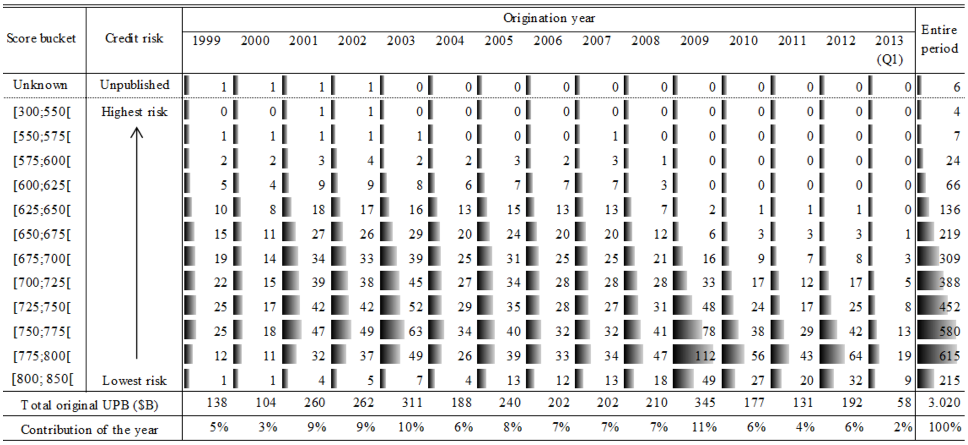
Borrowers’ score is a key indicator in mortgage lending. From 2009 onwards, the amount on loans in the scores bellow 625 is zero (Table 3.4), meaning that lending to low scored borrowers was firmly contained since then. By that year, lending moved markedly to the higher scores (see the shape of the bars moving between years 2008 and 2009, in Table 3.3 and Table 3.4). This effect is also captured in the score stability index that jumps from 0,10 to 0,28 in 2009 (Table 3.1, row 5). As a consequence, there is an increase in the concentration by scores from 14% in 2008 to around 20% in 2009 and in the following years (Table 3.1, row 4). Although this seems to be a reasonable prudential measure, we draw attention towards potential excessive lending bias and concentration in the highest scores, which requires a more precise risk-based pricing in these score bands. The number and amount of loans decreased after 2009.

Table 3.3: Freddie Mac database – number of loans, 1999-2013(Q1).



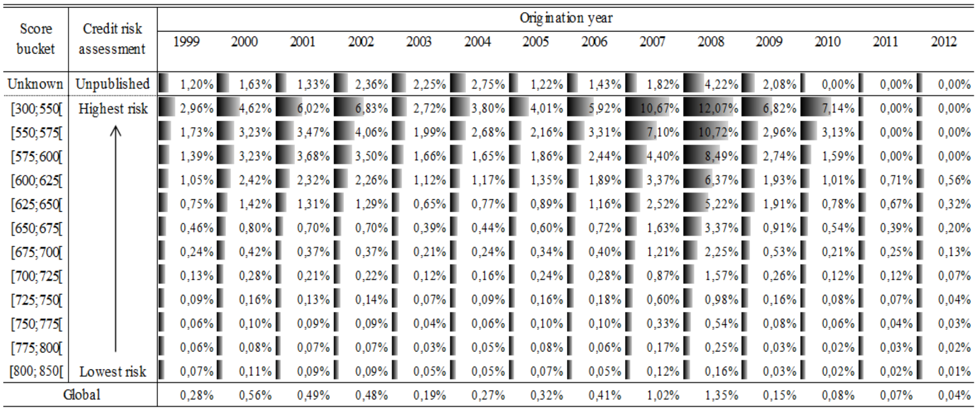
| Score bucket | Credit risk | Origination year | | | | | | | | | | | | | | | Entire period |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 (Q1) | |
| Unknown | Unpublished | 8 | 10 | 10 | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 40 |
| [300;550[ | Highest risk | 3 | 4 | 7 | 6 | 3 | 2 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 30 |
| [550;575[ | | 5 | 5 | 10 | 9 | 4 | 3 | 3 | 3 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 47 |
| [575;600[ | | 13 | 12 | 25 | 27 | 15 | 14 | 15 | 15 | 17 | 6 | 1 | 0 | 0 | 0 | 0 | 160 |
| [600;625[ | | 40 | 32 | 67 | 63 | 50 | 40 | 43 | 39 | 41 | 16 | 3 | 1 | 1 | 1 | 0 | 437 |
| [625;650[ | | 84 | 61 | 127 | 112 | 103 | 78 | 85 | 75 | 74 | 38 | 13 | 8 | 5 | 5 | 2 | 869 |
| [650;675[ | | 122 | 85 | 183 | 168 | 178 | 123 | 133 | 111 | 110 | 65 | 30 | 18 | 13 | 13 | 5 | 1.355 |
| [675;700[ | | 151 | 103 | 227 | 213 | 242 | 151 | 170 | 135 | 134 | 102 | 75 | 43 | 31 | 37 | 13 | 1.828 |
| [700;725[ | | 171 | 112 | 257 | 239 | 275 | 160 | 183 | 149 | 143 | 130 | 146 | 79 | 53 | 70 | 23 | 2.190 |
| [725;750[ | | 195 | 125 | 278 | 262 | 313 | 172 | 186 | 146 | 139 | 143 | 210 | 106 | 74 | 104 | 34 | 2.487 |
| [750;775[ | | 193 | 135 | 309 | 303 | 379 | 197 | 214 | 167 | 162 | 182 | 328 | 162 | 116 | 166 | 53 | 3.066 |
| [775;800[ | | 98 | 89 | 225 | 240 | 315 | 160 | 212 | 172 | 170 | 211 | 473 | 240 | 172 | 250 | 76 | 3.103 |
| [800; 850[ | Lowest risk | 11 | 13 | 32 | 37 | 52 | 30 | 77 | 71 | 74 | 91 | 235 | 130 | 91 | 142 | 41 | 1.125 |
| Global | | 1.095 | 787 | 1.757 | 1.685 | 1.930 | 1.131 | 1.324 | 1.083 | 1.069 | 986 | 1.513 | 788 | 556 | 787 | 247 | 16.737 |
| Contribution of the year | | 7% | 5% | 10% | 10% | 12% | 7% | 8% | 6% | 6% | 6% | 9% | 5% | 3% | 5% | 1% | 100% |

Table 3.4: Freddie Mac database – amount granted, 1999-2013(Q1). Unit: US B$.



| Score bucket | Credit risk | Origination year | | | | | | | | | | | | | | | Entire period |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 (Q1) | |
| Unknown | Unpublished | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| [300;550[ | Highest risk | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| [550;575[ | ↑ | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| [575;600[ | | 2 | 2 | 3 | 4 | 2 | 2 | 3 | 2 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 24 |
| [600;625[ | | 5 | 4 | 9 | 9 | 8 | 6 | 7 | 7 | 7 | 3 | 0 | 0 | 0 | 0 | 0 | 66 |
| [625;650[ | | 10 | 8 | 18 | 17 | 16 | 13 | 15 | 13 | 13 | 7 | 2 | 1 | 1 | 1 | 0 | 136 |
| [650;675[ | | 15 | 11 | 27 | 26 | 29 | 20 | 24 | 20 | 20 | 12 | 6 | 3 | 3 | 3 | 1 | 219 |
| [675;700[ | | 19 | 14 | 34 | 33 | 39 | 25 | 31 | 25 | 25 | 21 | 16 | 9 | 7 | 8 | 3 | 309 |
| [700;725[ | | 22 | 15 | 39 | 38 | 45 | 27 | 34 | 28 | 28 | 28 | 33 | 17 | 12 | 17 | 5 | 388 |
| [725;750[ | | 25 | 17 | 42 | 42 | 52 | 29 | 35 | 28 | 27 | 31 | 48 | 24 | 17 | 25 | 8 | 452 |
| [750;775[ | | 25 | 18 | 47 | 49 | 63 | 34 | 40 | 32 | 32 | 41 | 78 | 38 | 29 | 42 | 13 | 580 |
| [775;800[ | | 12 | 11 | 32 | 37 | 49 | 26 | 39 | 33 | 34 | 47 | 112 | 56 | 43 | 64 | 19 | 615 |
| [800; 850[ | Lowest risk | 1 | 1 | 4 | 5 | 7 | 4 | 13 | 12 | 13 | 18 | 49 | 27 | 20 | 32 | 9 | 215 |
| Total original UPB ($B) | | 138 | 104 | 260 | 262 | 311 | 188 | 240 | 202 | 202 | 210 | 345 | 177 | 131 | 192 | 58 | 3.020 |
| Contribution of the year | | 5% | 3% | 9% | 9% | 10% | 6% | 8% | 7% | 7% | 7% | 11% | 6% | 4% | 6% | 2% | 100% |

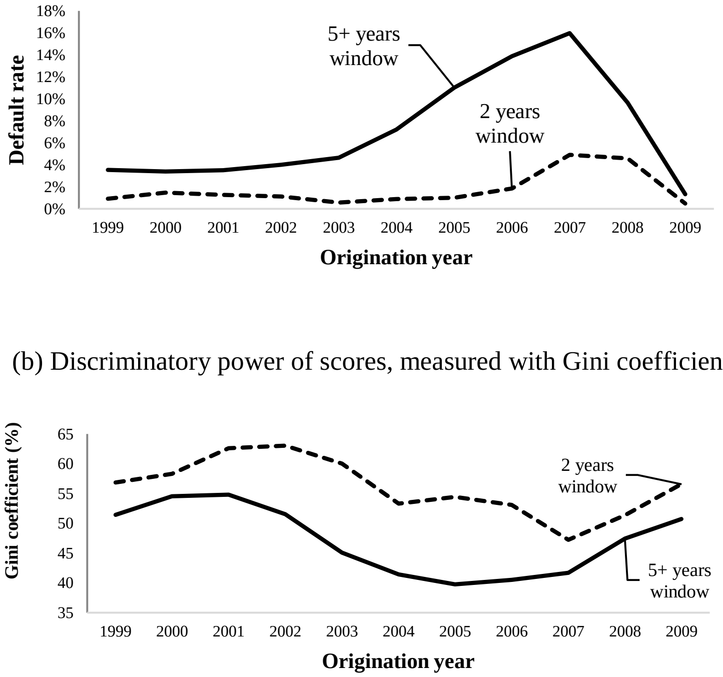
Table 3.5: Freddie Mac database – 1-year default by vintage, 1999-2013(Q1).



| Score bucket | Credit risk assessment | Origination year | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
| Unknown | Unpublished | 1,20% | 1,63% | 1,33% | 2,36% | 2,25% | 2,75% | 1,22% | 1,43% | 1,82% | 4,22% | 2,08% | 0,00% | 0,00% | 0,00% |
| [300;550[ | Highest risk | 2,96% | 4,62% | 6,02% | 6,83% | 2,72% | 3,80% | 4,01% | 5,92% | 10,67% | 12,07% | 6,82% | 7,14% | 0,00% | 0,00% |
| [550;575[ | ↑ | 1,73% | 3,23% | 3,47% | 4,06% | 1,99% | 2,68% | 2,16% | 3,31% | 7,10% | 10,72% | 2,96% | 3,13% | 0,00% | 0,00% |
| [575;600[ | | 1,39% | 3,23% | 3,68% | 3,50% | 1,66% | 1,65% | 1,86% | 2,44% | 4,40% | 8,49% | 2,74% | 1,59% | 0,00% | 0,00% |
| [600;625[ | | 1,05% | 2,42% | 2,32% | 2,26% | 1,12% | 1,17% | 1,35% | 1,89% | 3,37% | 6,37% | 1,93% | 1,01% | 0,71% | 0,56% |
| [625;650[ | | 0,75% | 1,42% | 1,31% | 1,29% | 0,65% | 0,77% | 0,89% | 1,16% | 2,52% | 5,22% | 1,91% | 0,78% | 0,67% | 0,32% |
| [650;675[ | | 0,46% | 0,80% | 0,70% | 0,70% | 0,39% | 0,44% | 0,60% | 0,72% | 1,63% | 3,37% | 0,91% | 0,54% | 0,39% | 0,20% |
| [675;700[ | | 0,24% | 0,42% | 0,37% | 0,37% | 0,21% | 0,24% | 0,34% | 0,40% | 1,21% | 2,25% | 0,53% | 0,21% | 0,25% | 0,13% |
| [700;725[ | | 0,13% | 0,28% | 0,21% | 0,22% | 0,12% | 0,16% | 0,24% | 0,28% | 0,87% | 1,57% | 0,26% | 0,12% | 0,12% | 0,07% |
| [725;750[ | | 0,09% | 0,16% | 0,13% | 0,14% | 0,07% | 0,09% | 0,16% | 0,18% | 0,60% | 0,98% | 0,16% | 0,08% | 0,07% | 0,04% |
| [750;775[ | | 0,06% | 0,10% | 0,09% | 0,09% | 0,04% | 0,06% | 0,10% | 0,10% | 0,33% | 0,54% | 0,08% | 0,06% | 0,04% | 0,03% |
| [775;800[ | | 0,06% | 0,08% | 0,07% | 0,07% | 0,03% | 0,05% | 0,08% | 0,06% | 0,17% | 0,25% | 0,03% | 0,02% | 0,03% | 0,02% |
| [800; 850[ | Lowest risk | 0,07% | 0,11% | 0,09% | 0,09% | 0,05% | 0,05% | 0,07% | 0,05% | 0,12% | 0,16% | 0,03% | 0,02% | 0,02% | 0,01% |
| Global | | 0,28% | 0,56% | 0,49% | 0,48% | 0,19% | 0,27% | 0,32% | 0,41% | 1,02% | 1,35% | 0,15% | 0,08% | 0,07% | 0,04% |

### 3.5.3. Risk over time – scores default misalignment

Fig. 3.4(a) shows the cumulative default for the aggregated loans in each origination year, and measuring the performance in two time windows: 2 years (dashed line) and 5 or more years after the loans have been underwritten (solid line). An interesting finding is that, although the mortgage bubble had expanded between 2001 and 2005, the higher default rates have occurred for the loans granted from 2004 through 2008. Loans originated in the beginning of the boom have not higher defaults than the loans in the previous years; only the borrowers that have underwritten after 2003 had defaulted more. This finding suggests that the first borrowers of the boom were just the “lucky ones” who borrowed cheaper; only then, the opportunistic (most likely fraud) and deluded borrowers entered in the “game”. Future research would much benefit from distinguishing default originated in fraud, possibly by depicting the early-stage delinquencies. Our results also reveal that, in relation to the year before crisis, in 2006, default rates more than doubled in the loans originated in 2007, from 0.41% to 1.02% and tripled by 2008, reaching 1.35% at the aggregate level (Table 3.5). This could be used as a benchmark for risk volatility when stress-testing current and future credit scoring and risk-based pricing models. Results also confirm that borrowers with the worst scores are more vulnerable to stressed conditions, e.g. unemployment and sudden credit-cuts, which intuition also suggests.



(a) Cumulative default rate in the loans' aggregate by origination year.

(b) Discriminatory power of scores, measured with Gini coefficient.

Fig. 3.4. Freddie Mac database - performance measures.

The top figure shows the cumulative default by origination year and the bottom figure shows the discriminatory power of scores. Measures are calculated in two windows: 2 years and 5+ years after the credit was originated.

In 2002, credit scoring models started to decrease their ability to discriminate between good and bad customers in the long-term (Fig. 3.4(b), solid line), and in 2003, they started to decrease in the short-term (Fig. 3.4(b), dashed line). This effect may be an outcome of population drifts, which requires further investigation of the influence of this event in credit assessment.

Finally, our research shows that real default rates by score are extremely irregular over time (Table 3.5), which requires further consideration in models alignment, either when they are used in credit decision making or in risk-based pricing designs.

## 3.6. Conclusions

Financial industry turned over-dependent on credit scoring in the advanced economies. A high proportion of the loan applications are automatically decided. In this framework, credit score is the central, if not the unique, indicator of the borrowers' credit risk.

We found evidence that the cumulative default by vintage has almost tripled in first years of crisis for scores equal to 650 or higher, suggesting that credit risk may has been under-priced in these cases. Two years after the crash, lending decision threshold changed and lending moved markedly to borrowers with higher scores, which led to an increase in concentration of lending in these individuals. Although this is a reasonable prudential measure, excessive lending bias and concentration towards the highest scores require more precise default estimation to correctly price credit risk.

Therefore, credit scoring models should properly adapt to time-changing conditions and lending dynamics, so that they faithfully support risk taking and pricing. Any misalignment between the PD's by score and the real default over time will guide to inconsistent decisions and suboptimal prices. There is a new emphasis on running predictive models with the ability of sensing themselves and learn adaptively (Gama, 2010, Adams et al., 2010, Pavlidis et al., 2012). This is one area where more sophistication is needed and more effort should be put to promote their wider acceptance.

# 4. A first approach to dynamic credit scoring

***Abstract* -** In this chapter we propose a two-stage model for dealing with the temporal degradation of credit scoring models. First, we develop a model from a classical framework, with a static supervised learning setting and binary output. Then, we introduce the time-changing economic factors, using a regression between the macroeconomic data and the internal default in the portfolio. In so doing, the specific risk is captured from the bank internal database, and the movement of systemic risk is determined with the regression. This methodology produced motivating results in a 1-year horizon, for a portfolio of customers with credit cards in a financial institution operating in Brazil. We anticipate that it can be extended to other applications of risk assessment with great success. This methodology can be further improved if more information about the economic cycles is integrated in the forecasting of default.

Parts of this chapter have been published in the proceedings of the conference BRICS Countries Congress on Computational Intelligence (BRICS-CCI), held in Brazil in 2013, with the title “A two-stage model for dealing with temporal degradation of credit scoring”. The methodology was used in the model than won the competition in the area of data mining and Finance promoted by that conference.

Some sections are repeated through the dissertation to make the chapters self-contained.

## 4.1. Introduction

In retail banking, credit risk assessment often relies in credit scoring models developed with supervised learning methods used to evaluate a person's credit worthiness, so-called scoring or PD models[1]. The output of these models is a score that translates a probability of a customer becoming a defaulter, usually in a fixed future period. Nowadays, these models are at the core of the banking business, because they are central to credit decision-making, in price settlement, and to determine the cost of capital. Moreover, central banks and international regulation have rapidly evolved to a structure where the use of these models is implicit, to achieve soundness standards for credit risk valuation in banking system.

Since 2004, with the implementation of regulations issued by the Basel Committee on Banking Supervision within Basel II, banks were encouraged to strengthen their internal models frameworks for reaching the A-IRB (Advanced Internal Rating Based) accreditation. To achieve this certification, banks had to demonstrate that they were capable of accurately evaluating their risks, complying with Basel II requirements, by using their internal risk models' systems, and keep their soundness. Banks owning A-IRB accreditation gained an advantage over the others, because they were allowed to use lower coefficients to weight the exposure of credit at risk, and benefit from lower capital requirements. A lot of improvements have been made in the existing rating frameworks, extending the use of data mining tools and artificial intelligence. Yet, this may have been bounded by a certain

[1] Other names can be used to refer to PD models, namely: credit scoring, credit risk models, scorecards, credit scorecards, rating systems or rating models, although some have different meanings.

unwillingness to accept less intuitive algorithms or models going beyond standard solutions being implemented in the banking industry, settled in-house or delivered through analytics providers (e.g. FICO, Experian, PwC and KPMG).

To our knowledge, developing and implementing a credit scoring model can be time and resources consuming – easily ranging from 9 to 18 months, from data extraction until deployment. Hence, banks use unchanged credit scoring models for several years. Bearing in mind that models are built using a sample file frequently comprising 2 or more years of historical data, in the best case scenario, data used in the models training are shifted 3 years away from the point they will be used. However, to our knowledge an 8-year shift is frequently exceeded. Should conditions remain unchanged, then this would not significantly affect the accuracy of the model, otherwise, its performance can greatly deteriorate over time. The recent financial crisis confirmed that financial environment significantly fluctuates unexpectedly, bringing renewed attention regarding scorecards built upon frames that are by far outdated. By 2007-2008, many financial institutions were using stale scorecards built with historical data of the early-decade. The degradation of stationary credit scoring models is an issue with empirical evidence in the literature (Crook et al., 1992, Lucas, 2004), however research is still lacking more realistic solutions.

Dominant approaches usually stand on static learning models. However, as the economic conditions evolve in the economic cycle, deteriorating or improving, the behaviour of the individuals also vary and their ability of repaying their debts. Furthermore, default evolution echoes trends of the business cycle, and related with this, regulatory movements, and interest rates fluctuations. In good times, banks and borrowers tend to be overoptimistic about the future, whilst in times of recession banks are swamped with defaulted loans, high provisions, and tighten capital buffers. The former leads to more liberal credit policies and lower credit standards, the later promotes sudden credit-cuts. Empirical evidence and theoretical frameworks support a positive, and lagged relationship between rapid credit growth and loan losses (Sousa et al., 2015a). Therefore, default needs to be regarded as time changing. Public studies are mostly motivated by capital consumption and regulatory concerns, and therefore they are focused in modelling macroeconomic factors. So far, none has explicitly integrated these factors in the existing credit scoring models.

Traditional systems that are one-shot, fixed memory-based, trained from fixed training sets are not prepared to process the highly detailed evolving data. Therefore, they are not able to continuously maintain an output model consistent with the actual state, or to quickly react to changes (Gama, 2010). These are some of the features of classic approaches that show that the existing credit scoring systems are limited. As the processes underlying credit risk are not strictly stationary, consumers' behaviour and default can change over time in unpredictable ways. There are several types of evolution inside a population, like population drifts, that translate into changes in the distributions of the variables, affecting the performance of the models. There is a new emphasis on running predictive models with the ability of sensing themselves and learn adaptively (Gama, 2010). Advances on the concepts for knowledge discovery from data streams suggest alternative perspectives to identify, understand and efficiently manage dynamics of behaviour in consumer credit in changing environments. In a world in which events are not preordained and little is certain, what we do in the present affects how events unfold in unexpected ways.

This chapter follows in section 4.2 with a brief description of the problem and the research objectives; in section 4.3 the database is succinctly presented. Section 4.4 details the methodology and theoretical framework of this chapter. A one-dimensional analysis is presented, as well as an overall assessment of the data available for modelling. A multidimensional approach is exposed in section 4.5 where we propose a two-stage model for dealing with the temporal degradation of credit scoring. In the first stage we develop a credit scoring, by comparing several supervised learning methods. In the second stage, we introduce the effect of time changing environment by shifting the initial predictions by a factor of the expected variation of default. Results in a 1-year horizon are presented in section 4.6. Conclusions and future applications of the two-stage model are discussed in section 4.7.

## 4.2. Problem and research objectives

The aim of this research is to propose a new approach for dealing with the temporal degradation of a portfolio of customers with credit cards in a financial institution operating in Brazil. Our work is attached to the BRICS 2013 competition, and is based in a real world dataset, along two years of operation, from 2009 to 2010. This competition consisted of two tasks, each focused on two features of the credit risk assessment model:

> *Task 1*: Develop a scorecard, tilting between the robustness in a static modelling sample and the performance degradation over time, potentially caused by market gradual changes along few years of business operation.

> *Task 2*: Fitting of the estimated delinquency produced by an estimation model to that observed on the actual data for the applications approved by the scorecard.

Participants were encouraged to use any modelling technique, under a temporal degradation or concept drift perspective. The official performance metrics were the area under the ROC curve (AUC) for task 1, and the Chi-square for the monthly estimates of delinquency in task 2. Innovative ways of handling task 1 can be found in PAKDD 2009 Competition whose focus was on this type of degradation. Task 2 represents an innovation in data mining competitions worldwide by emphasizing the quality of future delinquency estimation instead of the usual lowest future average delinquency.

We have built an integrated solution to deal with these two tasks. First we developed a credit scoring using a set of supervised learning methods. Then we calibrated the output, based on a projection of the evolution in the default. This forecast considered both the evolution of default and the evolution of exogenous data series, echoing potential changes in the population of the model, in the economy, or in the market. In so doing, resulting adjusted scores translate a combination of the customers' specific risk with systemic risk. As in many other applied financial studies, this research is bounded by some practical limitations, like systematic noise in the data and short time series. Our view is that the choice of the most appropriate methods for developing a credit scoring is context specific, and therefore, we present a technical approach driven by the specifics of the problem. We have also taken in consideration the extent of meaningful and reliable data that is actually available for modelling.

## 4.3. Database and development environment

The research summarized in this chapter was conducted in a real-life financial dataset, comprising 762,966 applications of credit cards, from a financial institution in Brazil. Data for modelling was provided along two years of operation, from 2009 to 2010. Each customer in the modelling dataset is assigned to a delinquency outcome - good or bad. In this problem, a person is assigned to the bad class if she had a payment in delay for 60 or more days, along the first year after the credit has been granted. The delinquency rate in the modelling dataset is 27.3%. Two additional datasets from the next year, 2011, were used to test the performance achieved in the static modelling sample, namely the leaderboard and the prediction dataset. The leaderboard contains a sample of 60,000 records that were obtained aggregating subsamples of 5,000 applications of each month in 2011. Although the default outcome was not available in the leaderboard dataset, the discriminatory power of the model given by the receiver operating characteristic curve (AUC) could be measured during the modelling stage. Competitors were allowed to make submissions of their solutions in the leaderboard, and for each the AUC and the distance D were measured online. The prediction dataset was used for the final performance evaluation in the competition. This dataset has 444,828 applications in 2011, for which the default outcome was not available at any stage of the modelling. A summary of the files is presented in Table 4.1.

Table 4.1: Case study Brazil - dataset summary.

| Dataset | Records | Period | Target | Delinquency (%) |
|---|---|---|---|---|
| Modelling | 762,966 | 2009-2010 | Labelled | 0.273 |
| Leaderbord | 60,000 | 2011 | Unlabelled | Unknown |
| Prediction | 444,828 | 2011 | Unlabelled | Unknown |

The full list of variables in the original dataset was downloaded from the BRICS 2013 official website. The dataset contains 39 variables, summarized in Table 4.2 and one binary target variable with values 1 identifying a record in the bad class and 0 for the good class.

Table 4.2: Case study Brazil - variables summary.

| Type | # | Information |
|---|---|---|
| Numerical | 6 | Age, monthly income, time at current address, time at current employer, number of dependents, and number of accounts in the bank. |
| Treated as nominal | 13 | Credit card bills due date, 1st to 4th zip digit codes, home (state, city, and neighbourhood), marital status, income proof type, long distance dialling code, occupation code, and type of home. |
| Binary | 16 | Address type proof, information of the mother and father's names, input from credit bureau, phone number, bills at the home address, previous credit experience, other credit cards, tax payer and national id, messaging phone number, immediate purchase, overdraft protection agreement, lives and work in the same state, lives and work in the same city, and gender. |
| Date | 1 | Application date. |
| ID | 3 | Customer, personal reference, and branch unique identifiers. |

## 4.4. Methodology and theoretical framework

This research evolves from a one-dimensional analysis, where we come across the financial outlook underlying the problem, to a multidimensional approach, where we gradually develop and experiment a new framework to model credit risk. The one-dimensional analysis was tailored to gain intuition on the default predictors and the main factors ruling the dynamics of default. The multidimensional approach is at the core of the work presented in this chapter and was held in two stages. In the first stage we developed a credit scoring model from a classical framework, with a static learning setting and binary output. In the second stage, we used a linear regression between exogenous data (fully listed in Table 4.4) and the internal default for adjusting the predictions.

In the first stage, we used the exogenous data series disclosed by the Central Bank of Brazil (Banco Central do Brasil, 2011) and the Brazilian Institute of Geography and Statistics[2] to evaluate the effect of time changing economics in the default evolution. We used quarterly series available from January 2004 through December 2011. Nevertheless, for determining the fitting between them and the internal default, only the period 2009 to 2010 was considered. We used the coefficient of determination, r-square, to evaluate the quality of fitting between the exogenous series and the default in the analysed portfolio, in a one-dimensional basis.

[2] Source: www.tradingeconomics.com.

### 4.4.1. Data analysis, cleansing and new characteristics

Some important aspects of the datasets were considered, because they can influence the performance in the unlabelled datasets. These aspects regard to:

*Great extent of zero or missing values* – In exception to the variables 'lives and works in the same state' and 'previous credit experience', binary fields have 95% to 100% concentrated in one of the values, which turn them practically unworkable. The same occurs for the numerical variables 'number of dependents' and 'number of accounts in the bank', both with more than 99% of zeroes. The remaining variables were reasonably or completely populated.

*Outliers and unreasonable values* – The variable age present 0.05% of applications assigned to customers with ages between 100 and 988 years. A small proportion of values out of the standard ranges are observable in the variables credit card bills due day, monthly income and time at current employer.

*Unreliable and informal information* – low reliability on socio-demographic data is amplified by specific conditions in the background of this problem. This type of scorecards is usually based in verbal information that the customer provides, and in most of the cases no certification is made available. In 85% of the applications, no certification for the income was provided, and 75% do not have proof for the address type. Customers have little or no concern to provide accurate information. The financial industry is aware of this kind of limitations. However, in highly competitive environments there is little chance to amend them, while staying in the business. Hence, other than regulatory imperatives, no player is able to efficiently overcome data limitations. As currently there are no such imperatives in Brazilian financial market, databases attached to this type of models are expected to keep lacking from reliability in the near future.

*Shifts on the distributions of modelling examples* – The most noticeable shift is in the variable monthly income. Values shift from one year to the other, which is an outcome of the increases in the minimum wages and in the inflation. During the analysed period, slight variations are also observable in the geographical variables, which are possibly related with the geographical expansion of the institution. In the remaining characteristics, the correlation between the frequency distributions of 2009 and 2010 range from 99% to 100%, suggesting a very stable pattern during the analysed period.

*Data cleansing and new characteristics* - We focused the data treatment on the characteristics that were reasonably or fully populated. Fields state, city, and neighbourhood contain free text, and were subjected to a manual cleansing. Classes with 100 or less records were assigned to a new class "Other". We could observe that there may be neighbourhoods with the same name in different cities; and hence we concatenated these new cleansed fields into a new characteristic. Taking into account that the shift in the variable monthly may be related to an increase of the minimum wages and inflation, a new characteristic was calculated by multiplying the inflation rate in the year by the variable monthly income.

*Data transformation* - Variables were transformed using the weights of evidence (WoE) in the complete modelling dataset, using the WoE

$$\mathrm{WoE_i} = \ln\left(\frac{\mathrm{n_{G_i}}/\mathrm{n_G}}{\mathrm{n_{B_i}}/\mathrm{n_B}}\right), \tag{1}$$

where $n_{G_i}$ and $n_{B_i}$ are respectively the number of good and the number of bad in the bin i, and $n_G$ and $n_B$ are respectively the total number of good and bads in the population sample. The larger the WoE the higher is the proportion of good customers in the bin. Numerical variables were firstly binned using SAS Enterprise Miner, and then manually adjusted to reflect the domain knowledge. In so doing we aim to achieve a set of characteristics less exposed to overfitting. Cases where the calculation of the WoE rendered impossible - one of the classes without examples - were given an average value. The same principle was applied to values out of the expected ranges (e.g. credit card bills due day higher than 31).

*One-dimensional analysis* - The strength of each potential characteristic is measured using the information value (IV) in the training dataset

$$\mathrm{IV} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left( \mathrm{n}_{\mathrm{G}_\mathrm{i}}/\mathrm{n}_\mathrm{G} - \mathrm{n}_{\mathrm{B}_\mathrm{i}}/\mathrm{n}_\mathrm{B} \right) \mathrm{WoE}_\mathrm{i}, \tag{2}$$

where n is the number of bins in the characteristic. The higher the IV is, the higher is the relative importance of the characteristic in a univariate basis. The higher is the IV the higher is the relative importance of the characteristic. In a one-dimensional basis, the most important characteristics are age, occupation, time at current employer, monthly income and marital status, with information values of 0.368, 0.352, 0.132, 0.117, and 0.116, respectively. Remaining characteristics have 0.084 or less.

*Interaction terms* - Using the odds in each attribute of the variables, we calculated new nonlinear characteristics using interaction terms between variables to model the joint effects. We tested six combinations, for which we present the IV in Table 4.3.

Table 4.3: Case study Brazil - interaction terms information value.

| Combination | IV |
|---|---|
| Age ∗ Income | 0.315 |
| Age ∗ Occupation | 0.009 |
| Income ∗ Marital status | 0.208 |
| Income ∗ Occupation | 0.334 |
| Income ∗ Proof of income | 0.123 |
| Age ∗ Income ∗ Occupation | 0.007 |

### 4.4.2. Modelling changing environment and time series analysis

The methodology presented in this chapter aims to propose an innovation in credit scoring modelling, by fitting the delinquency estimated with a scorecard, based on the movement of specific factors in the financial and economic environments. We based our empirical study in the analysis of exogenous time series, described in Table 4.4. Among the exogenous series that are available, we expect that this set includes the exogenous factors that have a major influence in the behaviour of the individuals and in credit cards repayments or delinquencies, by consequence.

Table 4.4: Case study Brazil - exogenous data series.

| Series | Correl[a] | r-square |
|---|---:|---:|
| Default on the financial revenue in credit cards in the country[b] | 0.805 | 0.648 |
| Default in revolving credit in the country[b] | 0.491 | 0.241 |
| GDP | -0.168 | 0.028 |
| GDP annual variation | 0.485 | 0.236 |
| Primary income payments (BoP, current US$) | 0.787 | 0.619 |
| Lending interest rate (%) | 0.118 | 0.014 |
| Real interest rate (%) | -0.047 | 0.002 |
| Taxes on income, profits and capital gains (% of revenue) | -0.256 | 0.065 |
| Household final consumption expenditure (% of GDP) | -0.713 | 0.509 |
| Private consumption | -0.365 | 0.133 |
| Inflation rate | 0.797 | 0.635 |
| Unemployment rate | -0.100 | 0.010 |
| Consumer confidence | -0.280 | 0.080 |
| Wages | 0.381 | 0.145 |

[a]. Correlation between the portfolio defaults rates series and the exogenous data series, from 2009 to 2010.

[b]. An abnormal observation was recorded in the first quarter of 2011. As we could not confirm the reliability of this value, we have chosen to replace it by the average of the adjacent quarters.

As most of the available exogenous data series are quarterly updated, we considered quarterly observations for all series. Although exogenous data series are available from 2004 onwards, we could not make a full use of them, because the internal data was only available for 2009 and 2010. Hence, we focused the analysis on that period, which may be considered short for achieving a fully reliable forecast. A minimum of 5 years is required with Basel II. In this type of analysis, it would be appropriate using a much larger historical period, to capture different phases of one or more economic cycles[3]. However, as the competition associated with this study aimed at an accurate 1-year forecast, our assumption was to consider that in 2011, Brazil would be in the same phase of the economic cycle as in the previous years - 2009 and 2010.

The internal default series follow very different paths in 2009 and 2010, Fig. 4.1, which discourages any attempt to discern an intra-annual seasonality. Nevertheless, in both years, the default slightly increased along the second semester. Although this is not a definite conclusion, we considered this occurrence in the forecasting scenarios, as we will describe hereafter.

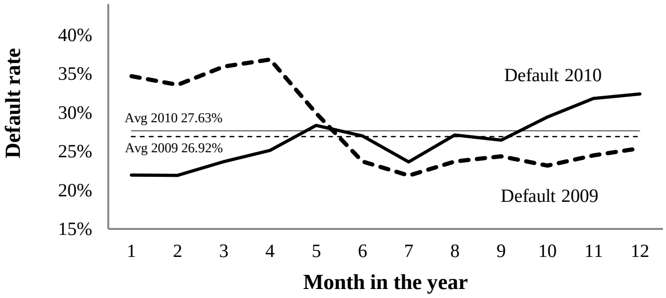


Fig. 4.1: Case study Brazil - internal default by month in 2009 and 2010.

In order to find potential relations between the internal default and the exogenous series, we calculated their correlations, Table 4.4. For the analysed period, the best series are the default on the financial revenues of credit cards in Brazil, primary income payments, households' final consumption

[3] An economic cycle may last for decades. Identifying an economic cycle is an important and non-trivial task that will be no further analysed here, as it goes far beyond the scope of this work.

expenditure, and the inflation rate. As only 8 observations are available, linear regression should consider a single independent variable, aiming to avoid overfitting. In forecasting scenarios, we only considered the series with the highest correlation - default on the financial revenues of credit cards in Brazil. Notwithstanding, an r-square of 64% in the regression can be considered low for the regression. We tested three forecasting scenarios, summarized in Table 4.5, which were iteratively submitted to the leaderboard. The final prediction is based on the scenario with the lowest distance D in the leaderboard.

Table 4.5: Case study Brazil - forecasting scenarios tested in task 2.

| # | Description | Rational |
|---|---|---|
| 1 | Estimate the default in each quarter adaptively from the values of default in credit cards of the previous quarter, and submit calculated values. | Incorporate new information adaptively, when it is available. This may benefit from the drift detection and suggest implementing corrective actions. |
| 2 | Estimate the default in each quarter adaptively from the values of default in credit cards of the previous quarter, and submit the average. | As there is no knowledge about the economic cycle, any correction resulting from a drift between quarters may be less apropriate in this application. It may be preferrable to use a central tendency of default (e.g.average defaults). |
| 3 | Submit the average annual default until September, and the average increased by 1% in the last two months of the year. | Use central tendency of default and adjust the months were the direction of the drift is more certain. |

## 4.5. Two-stage model

Some previous research suggests including economic conditions directly into a regression scorecard (Zandi, 1998), survival analysis (Bellotti and Crook, 2009), or transition models (Malik and Thomas, 2012). Our approach was to use a two-stage modelling framework in order to keep separate the two dimensions of risk – specific and systemic. The specific risk should be captured from the bank internal database with the scorecard developed in the first stage, and the movement of systemic risk is calculated with a linear regression at a second stage. The final score is therefore an adjustment of the initial score by the expected variation of the default in the population of the model.

### 4.5.1. Credit scoring construction

Several standard classification models, based on logistic regression (LR), AdaBoost, and Generalized Additive Models (GAM) were designed and tested with a 10 fold crossed-validation. Characteristics were iteratively added to the models, until no performance gain was observed in the test. Four different strategies were gauged, varying the window, and eliminating noise from the input dataset:

*High volume and diversity* - Use the entire modelling dataset (2009-2010), for more volume and diversity.

*Through-the-door* - Use the sample closest to the through-the-door population, for mitigating the effects of the temporal bias. We tested two different windows: 2010 full year, and 2010 last quarter.

*Ensemble of 12 models* - Create a model to apply in each month of the year, using the corresponding months in 2009 and 2011. Our idea is to address the changes in demand that may be attached to seasonal effects. Twelve models were developed, and their results were combined in the final score.

*Cleaning* - To overcome the presence of noise in the data, create a model using the entire modelling dataset (2009-2010); since the presence of noise in the data can spoil the model, we removed the examples strongly misclassified, and retrained the model in the reduced set. We consider an example strongly misclassified if the predicted posterior probability of the true class is less than 0.05.

As the financial institution is expanding, the leaderboard and the prediction sets should contain new codes in the discrete variables (e.g. new branch and geographical codes). These cases cannot be trained from the modelling dataset, because they cannot be observed before the expansion. As there was no knowledge about the nature and strategy of the expansion, we opted to assign an educated guess - use the average partial score for unfamiliar codes. In real-word environments, this assignment can be better guided, because the expansion strategy is known in advance.

### 4.5.2. Introducing the effect of time changing environment

In the second stage of modelling we introduced the effect of time changing environment into the scorecard previously developed. The initial predictions of the scorecard were shifted according to a factor of the expected variation of default. Every cut-off point set between 0 and 1 is converted into a new adjusted default rate within the set of approved applications, as illustrated in Fig. 4.2.

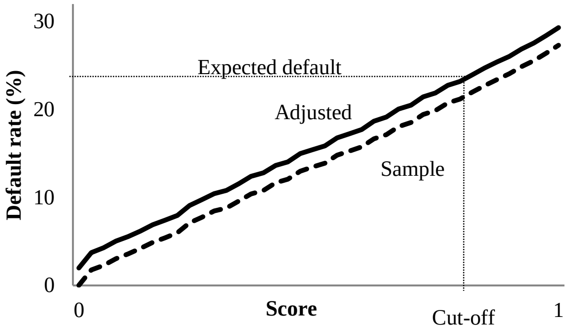


Fig. 4.2: Case study Brazil – model adjustment with the default central tendency.

## 4.6. Results

*Task 1* - Only the models with the best results in the test set were submitted to the leaderboard. Table 4.6 shows the results for the best models in the modelling dataset. Although a broad number of different configurations were tested, we only present the results for the models with an AUC higher than 0.71 in the test set. This includes the models that were also submitted to the leaderboard.

Table 4.6: Case study Brazil - best models in task 1, AUC>0.71 in the test set.

| Method | Period | Test | Leaderboard[a] |
|---|---|---|---|
| GAM | 2009-10 | 0.7320 | 0.7227 |
| GAM | 2010 | 0.7140 | 0.7131 |
| GAM | 2010(Q4) | 0.7204 | n.a. |
| LR | 2009-10 | 0.7222 | n.a. |
| LR with cleaning | 2009-10 | 0.7222 | n.a. |
| LR monthly | 2009-10, month against month | 0.7230 | 0.7140 |
| AdaBoost | 2009-10 | 0.7180 | n.a. |

[a.] n.a. - not available, because the model was not submitted to the leaderboard.

The model with the best results in the leaderboard is based on the Generalized Additive Model, and using the entire modelling dataset (2009 and 2010). This model was also submitted in the final prediction of the competition. Although we have tested different timeframes when modelling, it became apparent that for this application, the model with best performance was achieved with the larger volume of examples and diversity. The degradation of the proposed scorecard from the test (2009-2010) to the leaderboard (2011) is of 0.93%, which is quite acceptable in similar real-world applications. The degradation was partly controlled by adjusting unstable variables (e.g. monthly income based on the inflation rate). Apart from these, the bulk of remaining the characteristics is quite stable over time.

*Task 2* – The results of this task, presented in Table 4.7, demonstrate that when the forecast depend on very short time series, the best fitting is achieved with the simplest adjustment – the central tendency of default. In this problem, submitting an average default enhanced the fitting of default (scenario 2), which was further improved by adjusting on the months where the direction of the drift is more certain (scenario 3).

Table 4.7: Case study Brazil - results in task 2.

| **Scenario (#)** | **Average default rate** | **Distance D** |
|---|---|---|
| 1 | 0.293 | 3.16 |
| 2 | 0.293 | 1.45 |
| 3 | 0.294 | 0.83 |

[a.] Annual average forecasted default rate in 2011, if the entire portfolio is approved.

## 4.7. Conclusions

Theoretical models for knowledge extraction from data streams seem suitable for dealing with temporal degradation of credit scoring models. The idea is to use adaptive models, incorporating new information when it is available. Integrating new information may also benefit from the drift detection, and the occurrence of a drift may suggest eventual corrective actions to the model. Some specifics of the financial problems may turn the models quite stable along time, which is the case of the scorecard presented in this research. A static learning setting was at the basis of the model with the best discriminatory power. It also becomes apparent that some sort of time discretization may turn useless in some applications and may lead to nonsense or suboptimal forecasts. In this problem, as there is not an intra-annual seasonality of default, the practical meaning of a monthly prediction is may be debatable. Credit risk assessment is one area where the data mining and forecasting tools have largely expanded over the last years. However, there are a few areas where the use of these tools should focus essentially on providing a direction, rather than providing a strict prediction. There are a number of possible directions that no model that looks just into the past can enlighten about the future. This includes the directions driven by the business strategy of the bank (e.g. expanding the branch network, to offer a new product, or to merge with another financial institution). The same applies to the occurrence of extreme or rare events, like those that were roused by the recent financial crisis. The new paradigm of forecasting turns out to be looking at hidden streams in the present signal and understand how they will possibly direct an event into the future.

We propose a two-stage model for dealing with temporal degradation of credit scoring, which provided good results in a 1-year timeframe. However, it should also have a good performance in the long run. Therefore, future applications of this modelling framework should be tested in larger timeframes and consider lagged periods. Stress-testing this methodology should consider environments under major macroeconomic distress, or drifting populations resulting from expressive growth in the portfolios.

# 5. A new dynamic modelling framework for credit risk assessment

***Abstract*** - In this chapter we investigate the two mechanisms of memory, short-term (STM) and long-term memory (LTM), in the context of credit risk assessment. These components are fundamental to learning but are overlooked in credit risk modelling frameworks. As a consequence, current models are insensitive to changes, such as population drifts or periods of financial distress. We extend beyond the typical development of credit score modelling based in static learning settings to the use of dynamic learning frameworks. Exploring different amounts of memory enables a better adaptation of the model to the current states. This is particularly relevant during shocks, when limited memory is required for a rapid adjustment. At other times, a long memory is favoured. An empirical study relying on the Freddie Mac's database, with 16.7 million mortgage loans granted in the U.S. from 1999 to 2013, suggests using a dynamic modelling of STM and LTM components to optimize current rating frameworks.

Parts of this chapter are published in the Journal of Expert Systems with Applications in the paper titled "A new dynamic modeling framework for credit risk assessment". ISSN: 0957-4174; DOI:10.1016/j.eswa.2015.09.055.

A paper describing an application of the previous methodology to the Freddie Mac's loans database was submitted to the Journal of Risk Model Validation with the title "Dynamic credit score modelling with short-term and long-term memories: the case of Freddie Mac's database", and is under review.

Some sections are repeated through the dissertation to make the chapters self-contained.

## 5.1. Introduction

More than half a century has passed since credit scoring models have been introduced to credit risk assessment and corporate bankruptcy prediction (Harold Bierman and Hausman, 1970, Altman, 1968, Smith, 1964, Myers and Forgy, 1963). With today’s advanced economies, a high proportion of the loan applications are automatically decided upon using frameworks where the credit score is the central, if not the unique, indicator of the borrowers’ credit risk. In the United States (U.S.), the FICO score is an industry standard, claimed to be used in 90% of lending decisions, to determine how much money each individual can borrow and to set the interest rate for each loan. In the OECD countries, banks that have adopted the Internal Ratings Based (IRB) approach, in Basel II Accord (Bank for International Settlements, 2006, Bank for International Settlements, 2004), are using their own credit scoring models as the basis of the regulatory capital calculation.

A credit scoring model is meant to be an intelligent system. The output is a prediction about a given entity defaulting in a future period. In practice, one often uses a score that varies linearly in a positive range (e.g. FICO score varies in the range 300-850). In this arena, many frameworks, adaptations to real-life problems, and intertwining of base algorithms were, and continue to be, proposed in the literature, ranging from statistical approaches, to state-of-the-art machine learning algorithms, from parametric models to non-parametric procedures, see the papers of Jones et al. (2015) and Orth (2013). Typical credit scoring systems are developed from static datasets. Subject to context specifics, and provided that certain requirements of the methods are met, a timeframe for the development is delimited at some point in the past. By referring to historical examples within such a timeframe, the model is designed using a supervised learning approach. The resulting model is then used, possibly for several years, without further adaptation. As a consequence, traditional static credit scoring models are quite insensitive to changes within financial environments, like gradual or abrupt population changes caused by hidden transformations, or disturbances in periods of major financial distress. In line with this idea, Amato and Furnine (2004) found that ratings do not generally exhibit sensitivity to the business cycle.

To some extent, credit scoring models development still need to better mimic the human learning established on experience. There are two basic mechanisms of memory, short-term memory (STM) and long-term memory (LTM), which are fundamental components of human experience and cognition. The former is easy to set up but readily forgotten; the latter may take longer to set up but tends to be more durable (Baddeley, 2012). The aim of this study is to find a clearer understanding of which type of memory configuration for the learning of credit scoring systems enables a rapid adaptation to changes. Hence, our analysis is set on two research questions: Is recent information relevant for improving forecasting accuracy? Does older information always improve forecasting accuracy?

Consumers' behaviour and default change over time in unpredictable ways. There are several types of evolution inside a population, for example population changes, that translate into changes in the distributions of the variables, affecting the models. The behaviour of the individuals and their ability to repay their debts change when the conditions within the economic cycle evolve. In good times, banks and borrowers tend to be overoptimistic about the future, whilst in times of recession banks are swamped with defaulted loans, high provisions, and tightened capital buffers. The former leads to more liberal credit policies and lower credit standards, the latter promote sudden credit-cuts. Empirical evidence and theoretical frameworks support a positive, and lagged relationship between rapid credit growth and loan losses (Sousa et al., 2015a).

In order to adapt the models' output to changes over time, institutions should calibrate their scoring models according to the most recent information. There is a new emphasis on running predictive models with the ability of sensing themselves and learning adaptively (Gama et al., 2014). Advances on the concepts for knowledge discovery from data streams suggest new perspectives to identify, understand and efficiently manage dynamics of behaviour in consumer credit in changing environments. In a world where events are not preordained and little is certain, what we do in the present affects how events unfold, and they may do so in unexpected ways. New concepts for adapting to change and modelling the dynamics in populations have been proposed in credit score modelling (Adams et al., 2010, Pavlidis et al., 2012, Sousa et al., 2013). In this research, we apply a dynamic modelling framework for credit risk assessment, consisting of a sequential learning of the incoming new data. The driving idea mimics the principle of films, by composing the model from a sequence of snapshots rather than a single photograph. Two memory configurations are used: a STM and a LTM. The framework implements a component for adapting to drift, which is motivated by the original ideas of Widmer and Kubat (1996) and Klinkenberg (2004). The projected modelling framework is able to produce robust predictions not only in stable conditions but also in the presence of changes.

Renewed empirical credit risk measures are presented in this research using the Freddie Mac's single family mortgage loan-level database, first released in 2013. The database covers 16.7 million of fully amortized, 30-year fixed-rate mortgages, originated in the U.S. between 1999 and the first quarter of

2013. Based on historically observed delinquencies, the performance of the adaptive modelling is assessed in each memory configuration, and for a baseline static model developed with the data of the beginning of the period. We show that existing frameworks could be largely improved by including adaptive learning techniques. In such a setting, insight is provided into a multicomponent memory approach, consisting of a model combining a durable LTM component together with a temporary component, like STM (that in an extreme case can work as an episodic memory).

To the best of the authors' knowledge, the work most similar to ours is by Pavidlis, Tasoulis, Adams and Hand (2012) where an adaptive online algorithm is used in the classification of credit applications. It is based on the formulation of a criterion that enables a classifier to adapt to changes without completely disregarding all previous information. In the presence of population drift it is assumed that recent examples are more representative of the current classification than others in the distant past. Assorted experiments in artificial datasets exhibiting drift suggest that the method has the potential to yield significant performance improvement over standard approaches. However, an application of the method to a real-world dataset consisting of 92,258 UPL applications accepted between 1st January 1993 and 30th November 1997 in the United Kingdom, revealed that the model was unable to outperform a static classifier built with the data from the beginning of the period, 1993. The authors provide insufficient comments regarding this finding, regardless of the existence of population drift in the dataset, which had been documented in a previous study of Kelly, Hand and Adams (1999). Our research is the first to document the dominance of the adaptive over static modelling frameworks in a real-world relevant financial dataset, the Freddie Mac's database.

### 5.1.1. How does the industry currently handle credit scoring model maintenance?

Developing and implementing a credit scoring model can be time and resource consuming, easily taking from 9 to 18 months, from data extraction up to deployment. Not infrequently, banks use unchanged credit scoring models for several years. If conditions remain unchanged, then this does not significantly affect the accuracy of the models. Otherwise, the models' performance can greatly deteriorate over time. The recent financial crisis has drawn attention to models built on outdated timeframes. During the crisis, many financial institutions were using stale credit scoring models built with historical data from the first half of the decade; and many did not change their models in the aftermath of the crisis. The statistical deficiencies and degradation of stationary credit scoring models are issues widely documented in early literature (Eisenbeis, 1978) and backed up by empirical evidence (Sousa et al., 2015a, Rajan et al., 2015, Lucas, 2004, Avery et al., 2004).

Before the IRB approach had been introduced in the Basel II Accord, the financial industry had been less motivated to rebuild credit scoring models. At the time, financial institutions often outsourced model development to external parties, while assigning some internal staff to these activities. Changes to the models were rare, because they were expensive and time-consuming. Currently, many of the banks using the IRB approach have internalized this activity, because they are required to closely monitor the performance of the models and suitably respond to changes. Not infrequently, this requires multiple local adjustments to the models to improve their accuracy, which may be as costly and time-consuming as developing a new model. The European Banking Authority reports that models' adjustment or calibration has not a common practice amongst regulators. Many countries do not define any specific rules and when they do, these are usually not made public. Moreover, different countries favour different calibration choices (EBA, 2013b).

The huge advances in processing power and in storage capacity, together with the progress in streaming analytics, suggest increased practicality of adaptive modelling frameworks. However, some regulators are unlikely to approve models that change over time. So, under current circumstances, banks are likely to keep using a model as long as possible without further adaptation. This can be worrying, especially if the models' performance significantly declines during shocks. The impact of such degradation might be amplified because of other risk parameters, such as Loss Given Default (LGD), rising sharply, which pushes up the costs for misclassification errors. An insight into this effect is provided in a recent study of Sousa, Gama, and Brandão (2015b), where the disturbances in the return on lending in different scenarios of LGD, and of the default rates until maturity are measured.

This research provides new evidence on the significant degradation of credit scoring models based on static learning, broadly used among academics and practitioners. It is hoped that this research will provide useful guidance for future regulation in retail banking.

### 5.1.2. Structure of the chapter

Section 5.2 will provide a brief description of the settings and concepts of the supervised learning problem and score formulation. It will also present the fundamental ideas of adaptive learning. In section 5.3, we will present the conditions behind our case study, by providing an overview of Freddie Mac's database and the main dynamics over the period 1999-2013(Q1). Section 5.4 will present the adaptive modelling framework used in our experimental design. In section 5.5, we will compare the performance of the adaptive learning procedures with a baseline static model, and will compare the results of the STM with the LTM configuration. We draw conclusions in section 5.6.

## 5.2. Methods for adaptation

Traditional methods for building a credit scoring model consider a static learning setting. The model is trained using a predefined sample of past examples and then used to score new examples; actual or potential borrowers in the future. This is an offline learning procedure, because the whole training dataset must be available when the model is built. The model can be used for prediction only after having completed the training, and it will not be re-trained while in use, possibly for years, independently of changes in the surrounding environment. Alternatively, one might build a model that is updated continuously by incoming data.

The question remains whether it is best to have a long-term memory or to forget past events. On the one hand, a LTM might be desirable because it enhances the space of observed configurations. On the other hand, many of those configurations may no longer be relevant to the current situation. A rapid adaptation to change is achieved within a short window, because it reflects the current situation more accurately. However, the performance of models built upon shorter windows might decline in stable periods. In credit score modelling, this has been indirectly discussed by practitioners and researchers when trying to understand the pros and cons of using a through-the-cycle (TTC) or point-in-time (PIT) scheme to calibrate the output of the scorecards to the current phase of the economic cycle. For years a PIT scheme was the only option, because banks had insufficient data. Since the implementation of the Basel II Accord, banks are required to store the default data for a minimum of 7 years and consider a minimum of 5 years for calibrating the scorecards.

One of the most intuitive ideas to adjust to changes is to keep rebuilding the model from a window that moves over the latest batches and use this model for predicting on the immediate future. This idea assumes that the latest instances are the most relevant for prediction and that they contain the information of the current situation (Klinkenberg, 2004). The accumulation of batches of data, for example, annually, monthly, or daily, generates a flow of data for dynamic modelling.

An original idea of Widmer and Kubat (1996) uses a sliding window of fixed length with a first-in-first-out (FIFO) data processing structure. Each window may consist of a single batch or multiple sequential batches, instead of single instances. At each new time step, the model is updated in two stages. In the first stage, the model is rebuilt based on the training dataset of the most recent window. In the second stage, a forgetting process discards the data that moves out of the fixed-length window. Incremental algorithms (Widmer and Kubat, 1996) are a less extreme hybrid approach that allows for updating the models to the new context. They are able to process examples batch-by-batch, or one-by-one, and update the prediction model after each batch, or after each example. Incremental models may rely on random previous examples, or on representative selected sets of examples, called incremental algorithms with partial memory (Maloof and Michalski, 2004). The challenge is to select an appropriate window size.

## 5.3. Case study

Our research was conducted using the Freddie Mac's single family mortgage loan-level database, first published in March 2013. It tracks the performance of 16.7 million of fully amortized 30-year fixed-rate mortgages loans in the U.S., granted between January 1st 1999 and March 31st 2013. Sharing this data follows the direction of the regulator, the Federal Housing Finance Agency (FHFA), as part of a larger effort to increase transparency and promote risk sharing. The primary goal of making this data available was to help investors build more accurate credit performance models in support of the risk sharing initiatives highlighted by the FHFA in the 2013 conservatorship scorecard. The dataset is live data updated over time, typically at the end of each quarter, with the application and performance data being summarized by month, from the application point until the most recent reporting period.

### 5.3.1. Origination data

We considered a set of 16 variables that were available to the lenders at the time of the mortgage being granted, see Table 5.1. The release changes of the database are published online alongside a general user guide describing the full file layout and data dictionary (Freddie Mac, June 2013b). Freddie Mac's information regarding the key loan attributes and performance metrics can be linked to our research in the aggregated summary statistics (Freddie Mac, June 2014b).

Table 5.1: Case study U.S. - data available to the lenders at the loan application.

| Name | Short description | Type |
|---|---|---|
| **Credit score** | A number summarizing the borrower's creditworthiness at the time of the origination date. | Numeric |
| **First homebuyer flag** | Indicates whether the borrower is a first-time home buyer. | Binary |
| **Metropolitan area** | Identified with the metropolitan statistical area (MSA) or metropolitan division (MD) based on census data. | Treated as categorical |
| **Mortgage insurance percentage**<br>**(MI%)** | The percentage of loss coverage that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan, at the time of Freddie Mac's purchase.<br>For insured loans, the MI may vary between 1% and 55%. | Numeric |
| **Number of units** | Denotes whether the mortgage is a one-, two-, three-, or four-unit property. | Numeric |
| **Occupancy status** | Denotes whether the mortgage type is owner occupied, second home, or investment property. | Categorical |
| **Original loan to value (LTV)** | Original mortgage loan amount divided by the lesser of the mortgaged property's appraised value on the note date or its purchase price (in case of purchase or refinance mortgages). | Numeric |

| Name | Short description | Type |
|---|---|---|
| | Ratios falling outside the range 6% and 105%, are disclosed as unknown. | |
| **Original debt to income (DTI) ratio** | Debt to income ratio is based on the following calculation:<br>*Debt*: the sum of the borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the borrower is making, divided by;<br>*Income*: the total monthly income used to underwrite as of the date of the origination of the mortgage loan.<br>Ratios greater than 65% or unknown are passed as null values.<br>Note: The disclosure of the dataset is subject to the widely varying standards originators use to verify borrowers' assets and liabilities. | Numeric |
| **Original amount** | The UPB of the mortgage on the note date, rounded to the nearest $1,000. | Numeric |
| **Origination channel** | Indicates whether the channel at the origination of the mortgage is a retail lender, a broker or a correspondent. Situations where a third party origination is applicable but the seller did not specify the broker or correspondent are distinguished in the dataset. | Categorical |
| **Prepayment penalty mortgage (PPM)** | Indicates whether the mortgage is a PPM. A PPM is a mortgage with respect to which the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal. | Binary |
| **Property state** | A code identifying the state or territory within which the property securing the mortgage is located. | Categorical |
| **Property type** | Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single Family home. Situations where the property state is unknown can be recognized in the dataset. | Categorical |
| **Postal code** | The postal code for the location of the mortgaged property. | Treated as categorical |
| **Loan purpose** | Indicates whether the mortgage loan is a purchase mortgage, a cash-out refinance mortgage, or a no cash-out refinance mortgage. | Categorical |
| **Number of borrowers** | Identifies whether there is a single borrower or more who are obligated to repay the mortgage note secured by the mortgaged property. | Treated as categorical |

### 5.3.2. Performance data

Loan performance information is provided on a monthly basis and includes the monthly loan balance, delinquency status and information regarding early termination events: voluntary prepayments in full; 180 days delinquency (“D180”); repurchases prior to D180; third-party sales prior to D180; short sales prior to D180; deeds-in-lieu of foreclosure prior to D180; real estate owned (REO) acquisition prior to D180. Specific credit performance information in the dataset includes voluntary prepayments and loans that were short sales, deeds-in-lieu of foreclosure, third party sales, and REOs.

At the time of this research, data for performing loans and those that were up to 180 days delinquent was available through June 30th 2013. From the time it was granted until the most recent reporting period, there is a complete monthly historical report of the debt service for each loan, containing the following:

*Exposure at default value* - ending balance as reported by the servicer for the corresponding monthly reporting period.

*Loan delinquency status* - number of days that the borrower is delinquent, based on the due date of the last paid instalment reported by the servicers to Freddie Mac, calculated under the Mortgage Bankers Association (MBA) method. A code is used indicating the reason why the loan's balance was reduced to zero, in the following cases:

- Prepaid or matured (voluntary payoff);
- Foreclosed (short sale, third party sale, charge off or note sale);
- Repurchased prior to property disposition, or;

- Real-estate owned (REO) disposition.

We consider that a borrower defaulted if he was, at any point, 90 or more (90+) days delinquent, the typical definition used under Basel II. Later, in section 5.4, we will describe the construction of scorecards based on a supervised learning procedure and a binary target, where a borrower is assigned to the "bad" class, if he defaulted, and is assigned to the "good" class otherwise.

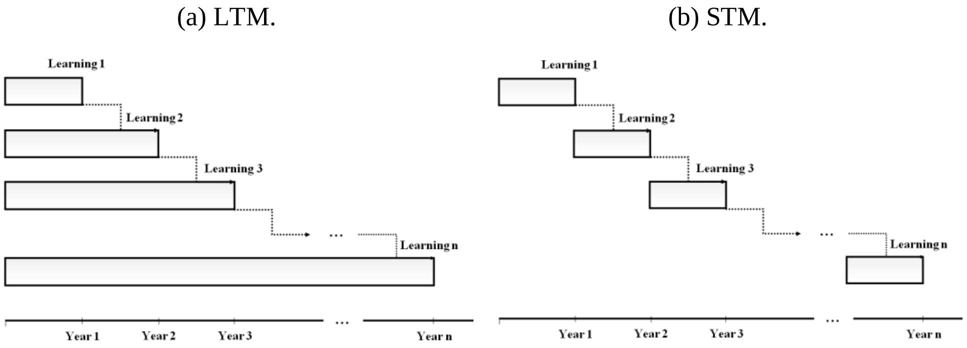
## 5.4. Modelling framework

### 5.4.1. Adaptive modelling

The dynamic modelling framework implemented in this research considers that data is processed batch-by-batch, as illustrated in Fig. 5.1. Sequentially, every year, a new model is built from a previously selected window, including the most recent year. To have sufficient performance window length we chose not to use loans granted from 2012 onwards.

In each model retraining - learning unit - we use a static setting. Each year, instances for modelling are selected from all previously available batches, according to a selection process. We use instance selection methods to test the hypothesis under investigation. Two methods were implemented – a LTM and a STM windowing configuration with a forgetting mechanism.

The LTM windowing configuration assumes that the learning algorithm generates the model based on all previous instances (Fig. 5.1(a)). The process is incremental, therefore every time a new instance arises, it is added to the training set, and a new model is built. This scheme should be appropriate to detect mild drifts, but it is unable to adapt rapidly to major changes. Models of this type should perform reasonably well in stable environments. A shortcoming of this incremental scheme is that the training dataset quickly expands, and this may require a huge storage capacity. In the STM windowing configuration, the model development uses the most recent window. With this scheme, Fig. 5.1(b), a new model is built in each new batch, by forgetting past examples. The fundamental assumption is that past examples have low correlation with the current default. Models of this type should quickly adapt to changes. A downfall of this method is that it often lacks the ability to generalize in stable conditions.


(a) LTM.

(b) STM.

Fig. 5.1. Case study U.S. - adaptive learning windowing configurations.

### 5.4.2. Constructing the scorecards

The classifier corresponding to each learning unit is a scorecard. Generalized Additive Models (GAM), introduced by Hastie and Tibshirani, are an extension of Generalized Linear Models (GLM) which, in turn, are an extension of Linear Regression (LR). Scorecards are GAMs, where the individual functions are piece-wise constant. The general approach to scorecard development involves the binning of the predictive variables and the optimization of the weight of each binned characteristic (Silva and Cardoso, 2015). A common practice is to compute the weights in two steps. Firstly, for each characteristic, the relative importance (score) of each bin is estimated; then, the relative importance of each characteristic is optimized. A standard way to estimate the relative importance of each bin is by using the weight of evidence (WoE) in the complete training dataset

$$\mathrm{WoE_i} = \ln\left(\frac{\mathrm{n_{G_i}/n_G}}{\mathrm{n_{B_i}/n_B}}\right), \tag{3}$$

where $n_{G_i}$ and $n_{B_i}$ are respectively the number of non-defaulted borrowers (good class) in the bin i and the number of defaulted borrowers (bad class) in the bin i, and $n_G$ and $n_B$ are respectively the total number of non-defaulted borrowers and total number of defaulted borrowers in the population sample. The larger the WoE is, the higher is the proportion of good borrowers in the bin. Numerical variables were firstly binned. Cases where the calculation of the WoE rendered impossible, i.e. no borrower following in one of the classes, are given an average value. The same rule is applied to values out of the expected ranges. The strength of each potential characteristic is measured using the information value (IV) in the training dataset

$$\mathrm{IV} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left( \mathrm{n}_{\mathrm{G_i}}/\mathrm{n}_{\mathrm{G}} - \mathrm{n}_{\mathrm{B_i}}/\mathrm{n}_{\mathrm{B}} \right) \mathrm{WoE}_{\mathrm{i}}, \tag{4}$$

where n is the number of bins in the characteristic. The higher the IV is, the higher is the relative importance of the characteristic in a univariate basis. Finally, the design of the scorecard is concluded by optimizing the weight of each characteristic using a linear model, as described in Silva and Cardoso (2015).

The scorecard design is wrapped in a forward feature selection process to find the optimal subset of characteristics. The selection process stops when no other characteristic adds significant contribution to the information value (IV) of the model. In this application the stopping criterion was set for a minimum increment of 0.03 in the IV. Tables 5.2 and 5.3 show the marginal contribution of the characteristics in each model adjustment, respectively, in the LTM and in the STM memory configurations. Cells are highlighted in grey if the characteristic was selected in the model adjustment. It’s worth noticing that, in the LTM configuration, the optimal subset of characteristics is more stable, and that the adjusted models tend to select a smaller number of characteristics. The STM configuration often leads to an adaptation based on a larger set of characteristics.

For the conclusions drawn from the experimental design to have validity, the same design process, as well as the same set of 16 potential predictors, was used in the learning units of both memory configurations (LTM and STM). In so doing, the difference in the performance of the models should be only due to the different time windows lengths.

Table 5.2: Case study U.S. - Marginal contribution in each LTM learning unit.

| Characteristic | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Credit score | 1,238 | 1,325 | 1,393 | 1,586 | 1,586 | 1,527 | 1,502 | 1,459 | 1,307 | 1,220 | 1,288 | 1,317 | 1,335 |
| First homebuyer flag | 0,000 | 0,001 | 0,006 | 0,002 | 0,002 | 0,000 | 0,000 | 0,000 | 0,001 | 0,002 | 0,003 | 0,003 | 0,003 |
| Metropolitan area | 0,080 | 0,071 | 0,056 | 0,043 | 0,043 | 0,028 | 0,029 | 0,026 | 0,028 | 0,021 | 0,021 | 0,022 | 0,021 |
| Mortgage insurance | 0,002 | 0,007 | 0,001 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,002 | 0,002 | 0,002 | 0,002 |
| Number of units | 0,006 | 0,000 | 0,001 | 0,000 | 0,000 | 0,000 | 0,001 | 0,001 | 0,001 | 0,003 | 0,004 | 0,003 | 0,003 |
| Occupancy status | 0,018 | 0,059 | 0,079 | 0,080 | 0,080 | 0,051 | 0,061 | 0,050 | 0,057 | 0,067 | 0,067 | 0,064 | 0,063 |
| Original debt to income | 0,051 | 0,016 | 0,020 | 0,014 | 0,014 | 0,022 | 0,021 | 0,024 | 0,039 | 0,069 | 0,077 | 0,084 | 0,088 |
| Original amount | 0,100 | 0,026 | 0,013 | 0,027 | 0,027 | 0,044 | 0,025 | 0,021 | 0,011 | 0,005 | 0,005 | 0,005 | 0,005 |
| Original loan to value | 0,159 | 0,130 | 0,203 | 0,245 | 0,245 | 0,264 | 0,236 | 0,226 | 0,247 | 0,259 | 0,261 | 0,258 | 0,259 |
| Origination channel | 0,085 | 0,064 | 0,090 | 0,102 | 0,102 | 0,093 | 0,075 | 0,073 | 0,080 | 0,166 | 0,121 | 0,102 | 0,095 |
| Prepayment penalty | 0,005 | 0,003 | 0,006 | 0,007 | 0,007 | 0,007 | 0,006 | 0,005 | 0,003 | 0,013 | 0,019 | 0,021 | 0,024 |
| Property state | 0,082 | 0,107 | 0,080 | 0,086 | 0,086 | 0,056 | 0,167 | 0,137 | 0,090 | 0,075 | 0,074 | 0,074 | 0,074 |
| Property type | 0,025 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,029 | 0,000 | 0,024 | 0,000 | 0,018 | 0,019 | 0,019 |
| Postal code | 0,039 | 0,064 | 0,029 | 0,022 | 0,022 | 0,019 | 0,017 | 0,018 | 0,016 | 0,016 | 0,018 | 0,018 | 0,019 |
| Loan purpose | 0,021 | 0,017 | 0,017 | 0,000 | 0,000 | 0,013 | 0,011 | 0,012 | 0,022 | 0,014 | 0,015 | 0,015 | 0,017 |
| Number of borrowers | 0,335 | 0,492 | 0,446 | 0,462 | 0,462 | 0,391 | 0,407 | 0,421 | 0,422 | 0,442 | 0,449 | 0,448 | 0,452 |
| Learning unit divergence | 2,245 | 2,381 | 2,440 | 2,676 | 2,676 | 2,515 | 2,586 | 2,474 | 2,350 | 2,374 | 2,440 | 2,453 | 2,478 |
| **Characteristics in the model** | **9** | **8** | **7** | **7** | **7** | **7** | **6** | **6** | **7** | **7** | **7** | **7** | **7** |

Table 5.3: Case study U.S. - Marginal contribution in each STM learning unit.

| Characteristic | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Credit score | 1,238 | 1,430 | 1,561 | 1,727 | 1,287 | 1,356 | 1,310 | 1,310 | 0,985 | 1,279 | 1,402 | 1,340 | 1,078 |
| First homebuyer flag | 0,000 | 0,007 | 0,008 | 0,008 | 0,001 | 0,000 | 0,001 | 0,001 | 0,000 | 0,000 | 0,000 | 0,000 | 0,026 |
| Metropolitan area | 0,080 | 0,022 | 0,046 | 0,046 | 0,072 | 0,070 | 0,088 | 0,088 | 0,029 | 0,025 | 0,207 | 0,254 | 0,365 |
| Mortgage insurance | 0,002 | 0,258 | 0,289 | 0,289 | 0,003 | 0,000 | 0,003 | 0,003 | 0,001 | 0,004 | 0,000 | 0,024 | 0,000 |
| Number of units | 0,006 | 0,001 | 0,001 | 0,001 | 0,008 | 0,001 | 0,003 | 0,003 | 0,004 | 0,004 | 0,004 | 0,000 | 0,000 |
| Occupancy status | 0,018 | 0,062 | 0,113 | 0,113 | 0,000 | 0,004 | 0,011 | 0,011 | 0,049 | 0,090 | 0,017 | 0,003 | 0,002 |
| Original debt to income | 0,051 | 0,012 | 0,026 | 0,014 | 0,041 | 0,043 | 0,022 | 0,022 | 0,042 | 0,055 | 0,250 | 0,264 | 0,347 |
| Original amount | 0,100 | 0,013 | 0,047 | 0,025 | 0,056 | 0,064 | 0,032 | 0,032 | 0,003 | 0,158 | 0,011 | 0,193 | 0,194 |
| Original loan to value | 0,159 | 0,008 | 0,013 | 0,013 | 0,304 | 0,247 | 0,119 | 0,119 | 0,404 | 0,322 | 0,224 | 0,003 | 0,318 |
| Origination channel | 0,085 | 0,085 | 0,174 | 0,174 | 0,072 | 0,040 | 0,015 | 0,015 | 0,076 | 0,058 | 0,043 | 0,020 | 0,015 |
| Prepayment penalty | 0,005 | 0,012 | 0,014 | 0,010 | 0,008 | 0,001 | 0,006 | 0,006 | 0,002 | 0,084 | 0,027 | 0,030 | 0,084 |
| Property state | 0,082 | 0,080 | 0,173 | 0,173 | 0,062 | 0,286 | 0,308 | 0,308 | 0,240 | 0,234 | 0,195 | 0,363 | 0,411 |
| Property type | 0,025 | 0,000 | 0,025 | 0,047 | 0,000 | 0,041 | 0,018 | 0,018 | 0,018 | 0,024 | 0,024 | 0,063 | 0,074 |
| Postal code | 0,039 | 0,022 | 0,014 | 0,026 | 0,068 | 0,053 | 0,027 | 0,027 | 0,026 | 0,020 | 0,065 | 0,162 | 0,160 |
| Loan purpose | 0,021 | 0,109 | 0,006 | 0,006 | 0,007 | 0,023 | 0,013 | 0,013 | 0,024 | 0,025 | 0,027 | 0,064 | 0,131 |
| Number of borrowers | 0,335 | 0,427 | 0,448 | 0,448 | 0,372 | 0,334 | 0,948 | 0,948 | 0,327 | 0,423 | 0,308 | 0,357 | 0,466 |
| Learning unit divergence | 2,2446 | 2,5486 | 2,9576 | 3,1196 | 2,3596 | 2,563 | 2,9224 | 2,9224 | 2,2312 | 2,8034 | 2,801 | 3,137 | 3,670 |
| **Characteristics in the model** | **9** | **7** | **8** | **8** | **9** | **10** | **6** | **6** | **7** | **9** | **8** | **10** | **11** |

The performance of the model is measured with the Gini coefficient, equivalent to the area under the ROC curve (AUC). It refers to the global quality of the credit scoring model, and may range between -1 and 1. The perfect scoring model fully distinguishes the two target classes, good and bad, and has a Gini index equal to 1. A model with a random output has a Gini coefficient equal to 0. If the coefficient is negative, then the scores have a reverse meaning. An extreme case of -1 would mean that all examples of the good class are being predicted as bad, and vice-versa. In this case, the perfect model can be achieved just by switching the prediction.

## 5.5. Results

We assessed the performance of the models sequentially learnt through the origination years 1999 to 2011. For each model rebuilding, the performance of the new model was measured in two sets: the modelling test set, containing a 20% random portion of the loans granted in the development year, and the set of loans granted in the following year, an out-of-sample performance.

The vintage curves presented in a previous study of Landy, Ashworth and Yang (2014) suggest that the cumulative default rates of this portfolio reached a plateau by the fifth year. Since most of the default events occurred between the first and the fifth year after the loan had been granted, we assumed that the performance measures of the models should be calculated within this timeframe. Therefore, despite the fact that the models' learning considered a fixed target concept - a borrower finding himself 90+ days delinquent at any point in a given timeframe after underwriting a loan -  performance was measured in five annually-incremental performance windows, from a 1-year to a 5-year performance window after the loan had been granted. In so doing, our aim is to bring awareness to the true performance of the models over the most relevant part of the life of the asset, rather than just interpreting the 1-

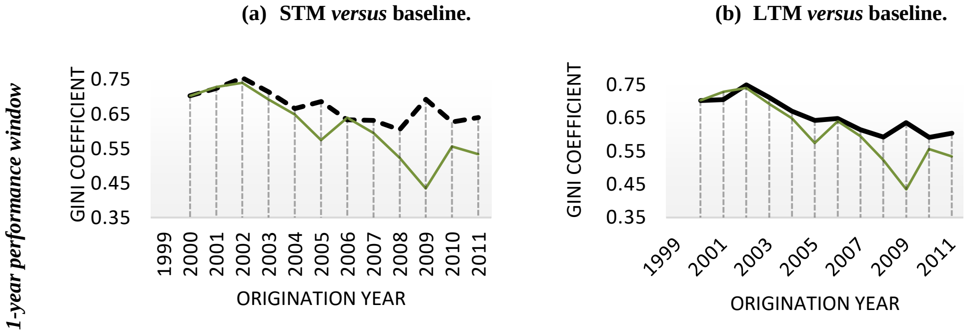
year performance window, as conventional approaches do. The last origination year for the performance measurements varies according to the length of the performance window (e.g. for the loans underwritten in 2009, only a 4-year performance window can be measured until 2013, and for the loans underwritten in 2012, only a 1-year performance window can be measured until 2013). Hence, the 5-year performance window is measured until the origination year 2008, the 4-year performance window is measured until 2009, the 3-year performance window until 2010 and the 1 and 2-year performance windows until 2011. The 1-year performance window is not presented for the loans granted in 2012, since the performance of the loans granted in December could only be measured through a half-year performance window, and this was deemed insufficient.

Below we will demonstrate the significant temporal degradation of static credit scoring in real-world environments, amplified during periods of major financial distress. Subsequently, we will present and discuss the results of the adaptive modelling framework, using the LTM and STM sliding-window configurations.

### 5.5.1. Adaptive learning versus baseline static learning model

A baseline static model was developed using the loans granted in the first year of the analysed period – 1999. This model was applied over the entire period, i.e. to each loan granted between 2000 and 2011, and the performance was assessed in each year, throughout the five performance windows. Results are presented in Fig. 5.2, where the performance of the adaptive learning models, in the STM and LTM configurations, is compared with the performance of the baseline static model. For a more realistic view, the results of the adaptive learning procedure consider that a model is applied to the loans granted in the year after the year used to train the model. In fact, a 2-year minimum window should be used to achieve a 1-year performance window for all the observations. We have chosen not to apply this principle due to the fact that we would have to disregard the performance for the year 2000 - the beginning of the housing bubble - that we are interested in. Considering the huge volume of available data, the learning could be based on a smaller sample (e.g. using a quarter instead of an entire origination year), which would allow an earlier readjustment of the model.

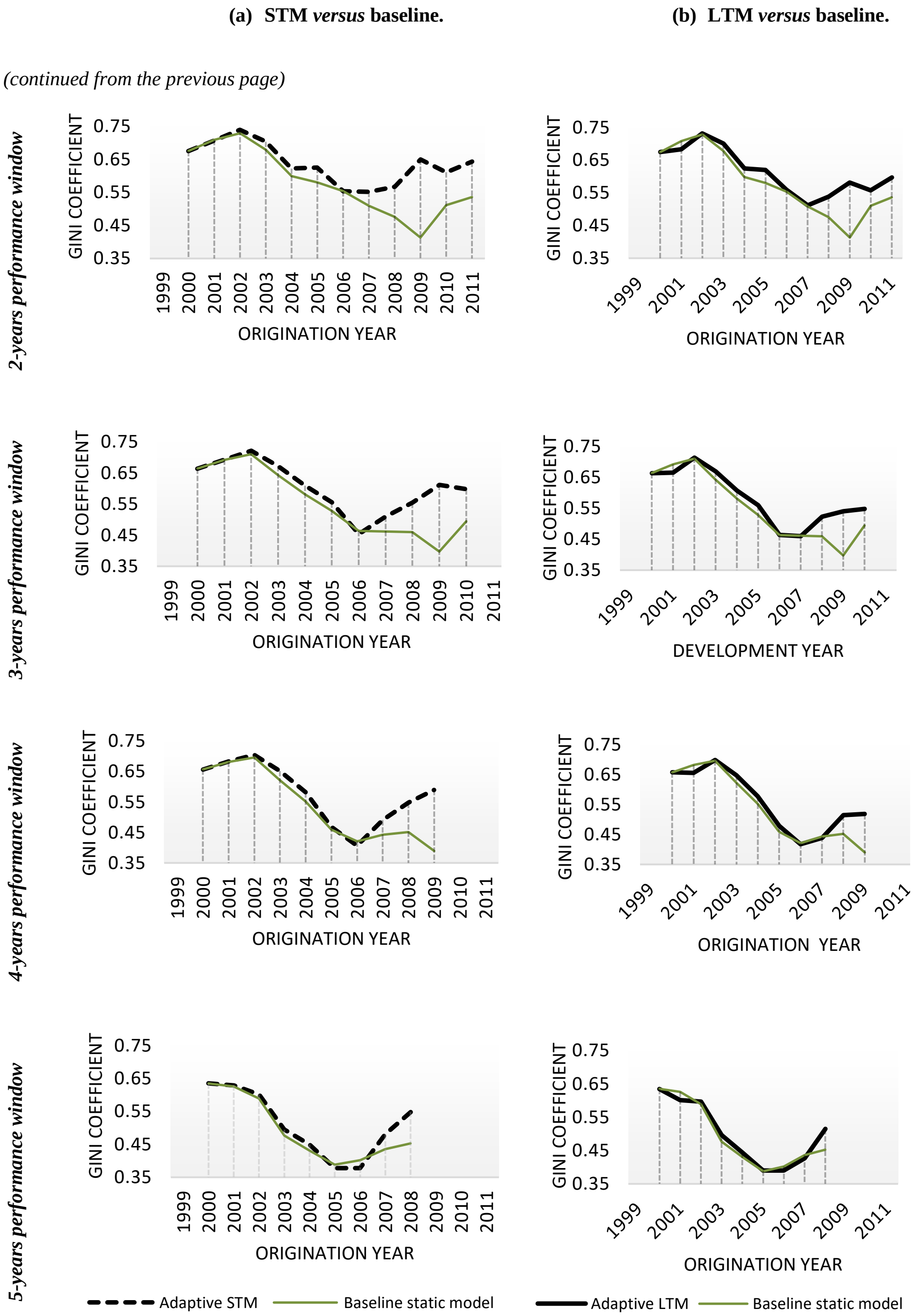
The performance of the baseline model gradually decreases over time, intuition also points to this. When compared with the adaptive learning procedure, the effectiveness of the performance decreases significantly from 2007 onwards and most noticeably in the aftermath of the crisis, in 2009. This finding is consistent for every performance window length.


**(a) STM *versus* baseline.**

**(b) LTM *versus* baseline.**

*(continues in the next page)*

(continued from the previous page)


**(a) STM *versus* baseline.**

**(b) LTM *versus* baseline.**

Fig. 5.2. Case study U.S. - adaptive learning *versus* baseline static model. Model applied to the loans originated 1 year after the development.

### 5.5.2. Adaptive short-term memory versus adaptive long-term memory

When comparing the performance of the short-term memory (STM) with the long-term memory (LTM) configuration in Fig. 5.3, we find that the STM configuration consistently outperforms the LTM. This finding is consistent both in the development test sample, referred to here as the development year, and in 1 year following the development. As it had been anticipated, the STM configuration consistently produced the highest performance during periods of exacerbated financial distress, from 2007 onwards. Even if we had speculated otherwise, the results of our analysis did not provide evidence that the LTM outperforms the STM in the analysed period. Our experimental design applies a LTM configuration that uses the longest available window until the point of relearning. However, this may not be sufficiently long to reveal a suitable range of memories and deliver dominant models in the LTM configuration. This is more likely to happen at the beginning of the period where the LTM configuration accumulates a few years' worth of history. We also speculated that the memory used in the STM configuration might still be too long, and that STM performance could have been further improved if we had tried shorter-term configurations. However, it is worth noticing that smaller windows may find it harder to gain approval by the industry, especially considering cost, business and regulatory constraints.

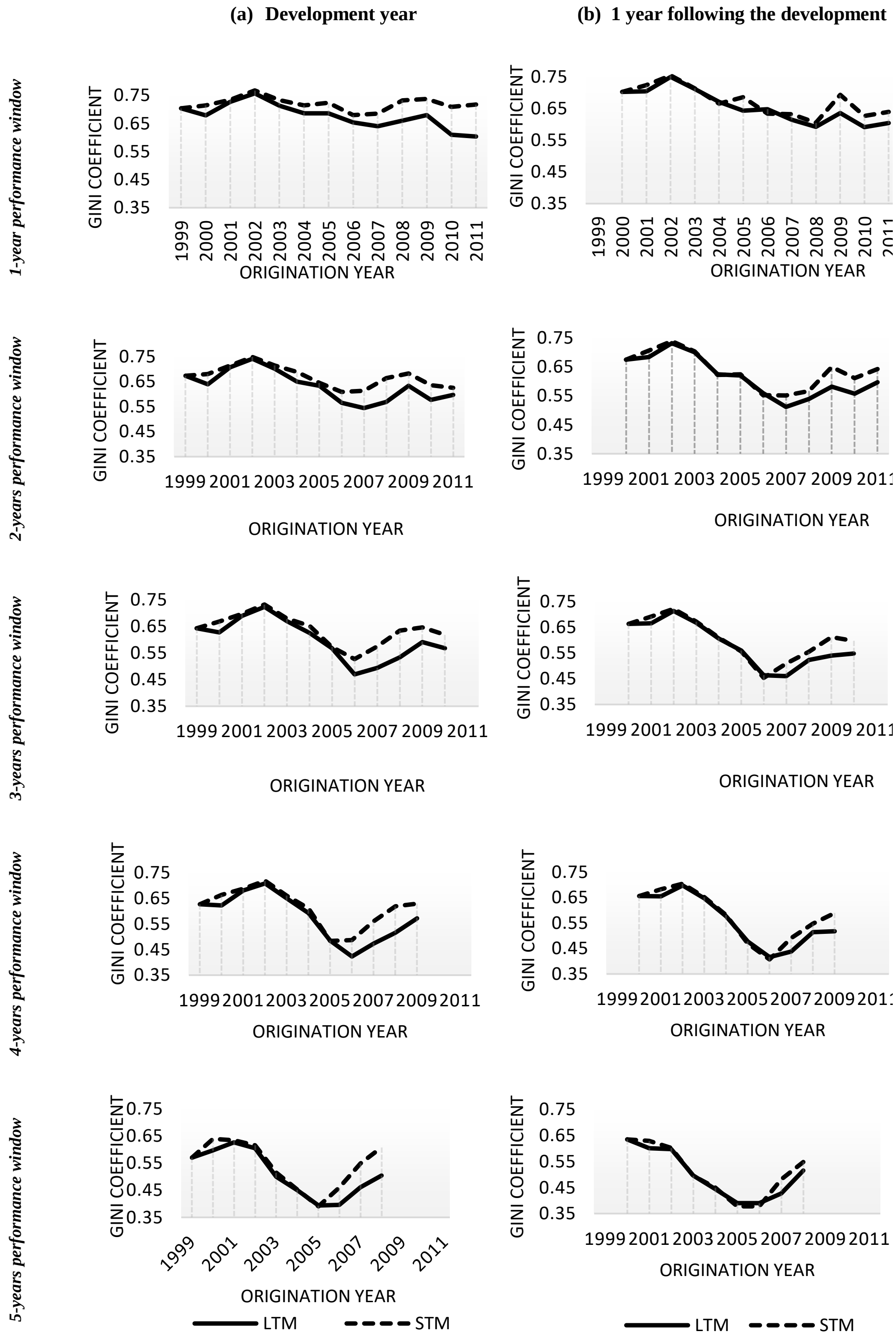


Fig. 5.3. Case study U.S. - performance of the adaptive learning.

## 5.6. Conclusions

Credit risk assessment is one area where data mining and forecasting tools have largely expanded over the last few years. In the advanced economies, credit scoring models are central to credit decision-making frameworks and to the contemporary internal rating systems since the Basel II Accord has been issued and implemented.

Typical credit scoring models are developed from static windows, and are therefore quite insensitive to changes, such as population changes or disturbances in periods of major financial distress. Theoretical models for knowledge extraction from data streams seem suitable for dealing with temporal degradation of credit scoring models. The idea is to use adaptive models, incorporating new information when it is available. Integrating new information may also benefit from detecting changes, and the occurrence of a change may point to eventual corrective actions applicable to the model. New concepts for adapting to changes have been proposed to deal with population drifts (Adams et al., 2010, Pavlidis et al., 2012, Sousa et al., 2013).

In this research we employ an adaptive modelling framework that stands on the original designs of Widmer and Kubat (1996) and Klinkenberg (2004). We are motivated to understand how the two basic mechanisms of memory, STM and LTM, influence the models' learning ability and predictive power through time. Central to our study is the idea that model learning is improved when mimicking human learning based on experience, and that STM and LTM are the driving components of that learning.

We present the performance of two types of adaptive modelling frameworks, STM and LTM. They were trained from a real-world dataset of 16.7 million loans that were at the epicentre of the global crisis, the Freddie Mac's single family mortgage loan-level dataset, first published in 2013. We did not attempt to challenge the existing adaptive modelling techniques. Instead, we aimed at using a straightforward adaptive learning framework to explicitly exhibit the STM and LTM capabilities in model learning. Two plain assumptions are confirmed in our investigation: newest data consistently improves forecasting accuracy, and STM allows a quick adaptation to changes. Older information did not improve forecasting accuracy, but no general rule can be made, since it may be an outcome of the context specifics. Although we had assumed otherwise, our empirical study did not reveal that the LTM outperforms the STM during stable phases. We speculate that this may have been a consequence of having used an insufficiently short window in the STM configuration. Our research presents renewed relevant empirical evidence that traditional modelling frameworks significantly degrade over time and that the models' predictive effectiveness is largely improved when adaptive learning frameworks are applied.

There are some real business problems with rebuilding models over time. Firstly, lenders have little incentive to enhance the existing rating systems' frameworks because it is expensive and time-consuming to build new scorecards. The scorecards need to be internally tested and validated, and then regulators need to approve them. Secondly, regulators still promote models whose coefficients do not change over time. This is one area where new evidence such as we have presented, might help. Our ideas for future work include trying to use ensembles of models that have been learnt from the past, instead of using the entire period to learn a new model. This has two major advantages. Firstly, a smaller sample is required for relearning the model, while still keeping memory from the past. Secondly, a model that depends on the previous assessments is more palatable; hence, it is more likely to be accepted. Another viable option is to develop a straightforward mechanism for modelling the link between the two components of memory identified in this study – LTM and STM. Regarding the STM, a prior selection of the window length seems appropriate and should be employed to optimize adaptation ability.

# 6. Risk management with dynamic defaults and under the Basel capital rules

***Abstract* -** Stress-testing is an important risk management tool used by banks and supervisors. The recent deployment of large databases enables the realistic evaluation of financial models. We present a first application of Freddie Mac's database recently released in 2013 to stress-testing, which can easily be replicated to other real-loans portfolios. Our proposal includes a first implementation of a return on risk-adjusted equity model embedded in the contemporary capital regulatory rules. Under this setting we analyse the impact of the Probability of Default (PD) and the Loss Given Default (LGD) in return on lending under the most extreme adverse circumstances.

Subprime lending was mostly banned from the primary markets. We claim that this is an overreaction to the global financial crisis. We find that much of the disturbances in the return on lending only occur for high LGD values. For LGD below 25%, simulations show that lending to borrowers with lower scores produces positive returns in the long-run. If sufficiently mature, these loans can boost portfolios' compositions, because they are less exposed to early repayments. So, rather than strictly declining these loans, regulators and lenders should ascertain the LGD boundaries under which the bank operates, to drive lending policies.

Parts of this chapter are published in the paper titled "Stress-testing the return on lending under real extreme adverse circumstances" which was presented at the EFMA Annual Meetings 2015. Taking in consideration the comments received, the paper was submitted with minor changes to the Journal of Credit Risk, with the title "Stress-testing the return on lending in the aftermath of the Crisis under the Basel Capital Rules", where it is under review.

Some sections are repeated through the dissertation to make the chapters self-contained.

## 6.1. Introduction

The subprime mortgage lending crisis in the U.S. came to public's attention when home foreclosures begun to rise in 2006 and moved out of control in 2007. A large decline in home prices prompted a devaluation housing-related securities and an unprecedented rise in mortgage delinquencies. Worldwide, banks' liquidity has plummeted with a significant disruption of the financing of businesses and consumers. This brought dramatic changes to financial regulation and banking supervision, including in the area of mortgages and consumer credit. Simultaneously, consumer financial protection regulation has been strengthened and expanded, and consumer financial behaviour has been changing so far.

Our study develops an empirical risk measure that relies heavily on the historical experience of delinquencies in this evolving landscape. In so doing, we contribute to enhanced knowledge in the area of credit risk assessment, which is relevant for setting regulatory and lending policies. We are focused on the study of the disturbances that affect the return on lending in real-world environments. The research question is if lending to the borrowers with the lower scores at the time of the loan application always produces negative returns in the long-run.

### 6.1.1. Stress-testing the return on lending under the Basel capital rules

Since 21 March 2013, Freddie Mac is making available loan-level credit performance data on a portion of fully amortized 30-year fixed-rate mortgages that the company purchased or guaranteed since 1999. The data is provided in a “living” dataset and by June 2014, the database covered over 16.7 million of fully amortized, 30-year fixed-rate mortgages in the U.S., originated between 1999 and the first quarter of 2013. Based on the historical observed delinquencies, we employ stress-testing as an attempt to project the returns under realistic extreme adverse economic scenarios. Returns are modelled under the current Basel capital rules, and so, the impact analysis is made considering two settings – applying the risk-weights as defined by the Basel Committee, and considering the parameters of the Final Rule under the Basel III for the U.S. banking. We stand on some basic ideas of  Saunders and Allen (2010) and on the model of Ruthenberg and Landskroner (2008), adapting the formulation of the model to our problem. We formulate the Basel Capital Requirements based on the work of Antão and Lacerda (2011), and in the Basel Committee regulation under Basel II and Basel III Accords (Bank for International Settlements, 2006, Bank for International Settlements, 2010).

### 6.1.2. Credit scoring – a central input in today’s lending

#### 6.1.2.1. Automated decision-making

Under the Basel capital rules, banks tailor their strategies to suitably remunerate their shareholders. In tailoring their activity strategies, commercial banks have two important decisions: selecting the borrowers’ desired risk profile, and positioning the pricing strategy, subjected to the market competitors and to regulatory boundaries.

When selecting the borrower’s desired risk profile, lenders determine if the risk of lending to a borrower is acceptable under certain parameters of credit risk, borrower’s credit capacity and collateral evaluation.

In retail lending, a great proportion of the loan applications are automatically evaluated. In this setting, credit score is the central, if not unique, indicator of the borrowers’ credit risk, either when the credit decision assessment is fully automatic or when it is an input for human decision. A person without a credit score or with a low score (meaning high risk) is unlikely to have credit, whilst an application of a person with a high score has good chances to be accepted.

An analysis on the causes and effects of the mortgage meltdown states that in 2007, 40% of all subprime loans have been generated by automatic underwritings in the U.S. This has been associated to lax controls in the underwriting processes (Bianco, 2008). Automated processes meant fasters decision, but less documentation scrutiny.

#### 6.1.2.2. Risk-based pricing

Banks increasingly use risk-based pricing models to price loans when positioning their pricing strategies, which are also moving towards credit score’s over-dependence. In the U.S., since its introduction 20 years ago, FICO score is calculated from the information available in the individuals’ credit bureau reports, and has become an industry standard. It is claimed to be used in 90% of lending decisions, to determine how much money each individual can borrow, and how much interest he will pay. As a result of the industry standards, the performance of credit loans depends on the credit scoring models accuracy, both in the short-term as in the long-run predictions. In 2007 and 2008, the delinquency rate in the mortgage loans in the U.S. rose sharply, both in borrowers in the lower scores as in the highest scores bands, showing that the actual risk of these borrowers has been underestimated.

### 6.1.3. The Freddie Mac's database – preliminary findings

Anderson, Scott, and Janet Jozwik (2014) proposed a framework for developing a credit model based on the Freddie Mac's dataset. For a 180-days delinquent target event, the authors conclude that much of the variation in credit performance across loans and over different stages of the economic cycle is explained by loan-level variables. Unsurprisingly, by adding factors to capture broader macroeconomic effects and the quality of underwriting, they significantly improve the model. Goodman, Landy, Ashworth and Yang (2014) present an exploratory paper providing a first look through the data, to find potential implications for guarantee pricing. The authors show the vintage composition as a percentage of the initial balance in a cross-analysis of the original borrowers' FICO score by the original loan to value (LTV). They follow the cumulative default in three groups in the score ranges 300 to 700, 700 to 750 and 750 to 850 crossed by the original LTV in selected buckets. The authors conclude that default rates are dramatically higher on higher LTV/lower scores, and so, they suggest that investors should look not only at the average LTV and FICO scores, but also at the FICO/LTV loans' distribution. The authors conjecture that pricing these portfolios by looking at averages should have led to under-priced default risk, but they do not present evidence. Sousa, Gama and Brandão (2015a) investigate the same database and find evidence that the first year cumulative default of the borrowers with score 650 or higher has almost tripled in first years of the last global crisis, suggesting that credit risk may has been under-priced in these cases. They show that default rates increased sharply during the crisis, but it did not increase uniformly along the range of borrowers' credit scores. They also show that, two years after the crash, lending decision threshold moved markedly to borrowers with scores higher than 625, which led to an increase in concentration of lending in the individuals in the highest score bands. Although this is a reasonable prudential measure, excessive lending bias and concentration towards the highest scores require more precise default estimation to correctly price credit risk.

### 6.1.4. Contributions of this study

Discussion is being pushed towards risk-based pricing designs under the current capital rules. Previous studies (Anderson and Jozwik, 2014, Goodman et al., 2014, Sousa et al., 2015a) suggest that risk-based pricing models used within today's lending are highly dependent on credit scores, which in turn are increasingly ruled under the Basel capital regulation. Behn, Haselmann and Wachtel (2015) use a quasi-experimental research design to examine the effect of model-based capital regulation on the pro-cyclicality of bank lending and firms' access to funds.

Complementing the previous works, we focus on the performance of returns in different arrangements of borrowers' credit risks, through contrasting phases of the economic and business cycles, including the years of exacerbated financial distress of the last global crisis.

This chapter follows in section 6.2 with a description of the return on risk-adjusted equity model that is used in our study. We describe the regulatory environment, under the current capital regulation, and the base concepts that rule the model. In section 6.3, we describe the conditions of our empirical study. Firstly, we make an overview of the database used in our experimental study. Secondly, we provide a descriptive analysis regarding the interest rates setting, and about the lending evolution over the analysed period. Finally, we describe the main assumptions of the stress-testing exercise of this research. Section 6.4 provides the results for nine tested scenarios. We depict a pessimistic scenario that has been drawn from the real historical default rates of mortgage loans in the U.S. during the global crisis. Conclusions are presented in section 6.5.

## 6.2. Return on risk-adjusted equity model

Our model assumes a commercial bank that operates in the primary market, raising deposits from the clients and extending credit to the public. It also operates in the secondary market, where it transacts with other commercial banks, with the central bank, and in the financial markets. The bank holds regulatory capital as defined under the internal ratings-based (IRB) issued in Basel II as a cushion against unexpected losses. We assume that the bank uses a risk-based pricing (RBP) model to price loans. The objective function is to maximize the expected profit as a function of credit risk, based on the decision variable, customer score, $s(\mathbf{x})$, which translate the probability of the customer entering in default in the loan after the credit has been granted.

### 6.2.1. Regulatory environment and base concepts

#### 6.2.1.1. Regulatory capital for credit risk – IRB setting

The Basel II Accord, established in 2004 and revised in 2006, attempted to implement more risk-sensitive credit exposure measures into capital requirements (Bank for International Settlements, 2006). Banks[4] were allowed to choose the way of determining the requirements of minimum capital, by selecting the methodology of calculating the risk-weighted approach. The Standardized approach is based on external credit risk assessments, while in the IRB, financial institutions use their internal credit risk models' system to determine the credit risk of each activity, such as commercial or consumer lending.

#### 6.2.1.2. Regulatory capital for credit risk – impact in lending pricing

There has been a great deal of talk about Basel II leading to greater risk-based pricing in loan markets, because the new rules for the risk-weighted assets amplified the difference between capital required for risky and safe lending

[4] Through this chapter the term "bank" is used to mean bank, banking group or other entity (e.g. holding company) whose capital is measured under the Basel Accord.

categories and borrowers. Lenders using the IRB, the Advanced banks, have a much lower cost of funding when lending safer types of debt or when lending to safer borrowers. This should have pushed lending away from riskier types of debt, and shorten the prices to safer categories such as low LTV mortgages.

The impact of the new rules on borrowers is hard to discern, because it is not certain that banks had the right incentives to affect the prices by the differences in the cost of capital according to the type of debt or risk of the borrower. In fact, many lenders were already using risk-based pricing, especially for higher risk lending such as subprime mortgages and consumer loans, to compensate the exposures' expected losses. In the one extreme, banks may not be motivated to reduce prices where the cost of funding got lower, in particular to safer borrowers, because their return on equity goes higher. In the other extreme, banks may have not been able to reflect the higher cost of funding to higher risks loans. The first reason is because many jurisdictions have issued consumers' protection laws that impose a maximum cap in the loans' rate. The second reason is because, worldwide, Advanced banks are playing in markets where there are competitors using the Standardized approach, which use a lower risk-weights constant parameter in the highest risk borrowers.

#### 6.2.1.3. Minimum capital requirements

In the aftermath of the global crisis, the Basel III capital framework that was agreed upon internationally in December 2010, and revised in June 2011, established the new minimum risk-based capital ratios to be adopted worldwide (Bank for International Settlements, 2010). The international package includes a new 4.5% common equity Tier I capital requirement, a 6% Tier I capital requirement, and retained the general requirement for banks to hold a minimum total capital, or regulatory capital, of 8% of their total risk-weighted assets (RWA), i.e.:

$$\frac{\text{Eligible regulatory capital}}{\text{Total RWA}} \geq 8\% \cdot \tag{1}$$

In addition to the minimum capital ratios, banks are required to maintain a capital conservation buffer of 2.5% of risk-weighted assets to avoid restrictions on their ability to distribute capital and to pay some discretionary bonuses payments to executive officers. Hence, the minimum capital ratios effectively increased to 7%, 8.5%, and 10.5%, respectively. Banks falling within the buffer will be required to limit dividends, share repurchases or redemptions, and discretionary bonuses. For banks using the IRB, the capital buffer may be increased during periods of extensive credit growth by an incremental countercyclical capital buffer of up to 2.5% of the risk-weighted assets.

Currently, for the exposures not in default, financial institutions using the IRB compute the total RWA by multiplying the capital requirements for market risk, $CR_M$, and operational risk, $CR_O$, by 12.5 (i.e. the reciprocal of the minimum capital ratio of 8%) and adding the resulting figures to the sum of risk-weighted assets for credit risk. Then a scaling factor of 1.06 is applied to the risk weighted assets for credit risk aiming to broadly maintain the aggregate level of minimum capital requirements (Bank for International Settlements, 2006).

$$\text{Total RWA} = 12.5(1.06 \times K \times EAD + CR_M + \ CR_O) \qquad (2)$$

where K is the capital requirement for the credit risk asset amounts, whose formula is disclosed within the Basel II official documents (Bank for International Settlements, 2006), and EAD is the exposure at default. According to a recent report of the European Banking Authority (EBA, 2013a) the RWA component related to credit risk for the aggregate of the European banks operating under the IRB represents about 77% of the total RWA.

The derivation of risk-weighted assets depends on the estimates of the PD[5], LGD, EAD and, in some cases, effective maturity (M), for a given exposure. The PD is a measure of the borrower's risk, the loss given default, LGD, is the expected proportion of the exposure that the financial institution will recover conditional to the borrower entering in default. The formulas for computing the K parameter vary according to the risk category of the assets, whether the exposure at risk belongs to one of the macro-segments: corporate, sovereign, banks and retail. Corporate segment includes the exposures of all enterprises, excepting the non-financial enterprises of small and medium size, SME, with exposures bellow 1 million Euros. Enterprises excluded from the corporate segment are considered in retail. In the retail segment, the risk weights are differentiated by the type of credit, whether it is in the category of residential exposure, a qualifying revolving credit line, or other retail exposure.

[5] The Basel Accords official documents typically use the abbreviation PD to denote the probability of a borrower entering in default, which is equivalent to the notation *p(B)*, probability of a borrower being bad, used in chapter 2. In the current chapter we will use the abbreviation PD to be consistent with the Basel terminology.

#### 6.2.1.4. Exposure at default

Our model considers a loan with the original conditions, at the time of the application: amount $A$, annual interest rate $R$, term of $n$ years and regular monthly payments. These assumptions determine the number of payments, $N = 12n$, the monthly interest rate, $r = R/12$, and installment amount:

$$M = \frac{rA}{1-(1+r)^{-N}} \tag{3}$$

A basic assumption is that the monthly instalment consists of the sum of two parts. One part consists of the interests' amount to be paid in the month and the other is the repayment of the initial amount granted, such that in the month i, i=1.. N, these two parts are computed as follows:

- Repayment of amount granted ($P_i$):

$$P_i = (M - rA)(1+r)^{i-1} \tag{4}$$

- Interests' amount to be paid ($I_i$):

$$I_i = M - P_i = rC_{i-1} \tag{5}$$

where $C_{i-1}$ is the exposure at default in the end of the month $i-1$ and in the beginning of the month $i$, i.e. the amount after amortization, which is computed as:

$$C_{i-1} = A - (M - rA)\sum_{k=1}^{i-1}(1+r)^{k-1} \text{ and } C_0 = A. \tag{6}$$

Equivalent to

$$C_{i-1} = A - (M - rA)\left(\frac{1-(1+r)^{i-1}}{r}\right) \text{ and } C_0 = A. \tag{7}$$

Here we assume that until the default event, the borrower entirely pays the instalments according to the debt service plan. For simplicity, in the current setting of model we do not distinguish the early repayments, which also impact on the expected returns. Therefore, the exposure at default in the month $i$, i=1..N, for a borrower that did not reach the default event is:

$$EAD_i = C_{i-1}. \tag{8}$$

#### 6.2.1.5. Cost of funding

An asset of amount $A$ is funded by equity, $E$, and debt, D, such that:

$$A = E + D. \tag{9}$$

The equity needed to fund the asset, related to credit risk, can be assumed as:

$$\mathrm{E} = sr \times rw \times \mathrm{A}, \tag{10}$$

where $rw$ $i$s the risk-weight factor for credit risk, and $sr$ is the solvability ratio targeted by the bank, which needs to be at least 8% - the minimum capital ratio. The $rw$ parameter is defined by the Basel Committee, both for the Advanced and for the Standardized approach, and may be adjusted by local bank regulatory agencies, as it happens in the U.S. and in the E.U. Examples of local conventions include the Final Rule issued by the U.S. bank regulatory agencies that set comprehensive regulatory capital framework for the U.S. banking organizations under the Basel III and implements the capital-related provisions of the Dodd-Frank Act. In the E.U., Basel Accords have been introduced via the Capital Requirements Directive (CRD). This Directive provides a common framework for implementation, but allows for national discretions. One example of a national discretion is the LTV limit for qualifying residential mortgages with the preferential risk weighting of 35%. In the UK, for example, this limit has been set at 80% by the Financial Services Authority (FSA). For a bank using the IRB approach without any changes rw is 12.5 × 1.06 × K.

Equation (9) can be rewritten as:

$$A = sr \times rwA + (1 - sr \times rw)A \tag{11}$$

The cost of equity, $C_E$, can be evaluated according to the return required by the shareholders, for which we assume the profitability measure return on equity, $r_E$.

$$C_E = r_E E. \tag{12}$$

Equity is the costliest way of financing. So, financial institutions using the IRB benefit from lower costs of funding when lending to entities of better risks, i.e. lower PD and lower LGD. In the other extreme their cost of funding is higher when lending to entities of worst risks, i.e. higher PD and higher LGD. This can easily be recognised from the monotonic behaviour of the $rw$ factor when varying these two parameters, which we illustrate with an example in Fig 6.1.

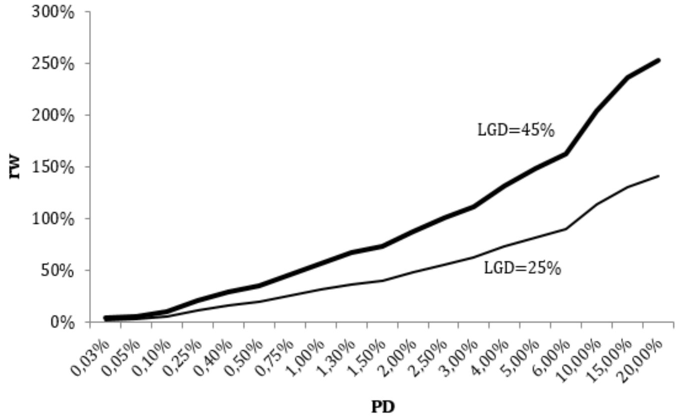


Fig. 6.1: Illustrative IRB risk-weights for residential retail.
Source: Bank for International Settlements (2006).

On the other side, the cost of debt, $C_D$, is

$$C_D = r_D D, \tag{13}$$

where $r_D$ is the price of debt, considering the structure of financing and price that the bank can get in wholesale funding (central banks and markets) and from clients' deposits.

### 6.2.1.6. Expected loss

Our model assumes that when a borrower enters in default then from that point onwards he will not leave the default status. If the $PD_i$, i=1.. N, is the probability of the borrower entering in default in the month $i$ and the LGD is the loss given default conditioned on the default event, then the expected loss for that borrower in that month, $EL_i$, is

$$EL_i = PD_i \times LGD \times EAD_i \qquad (14)$$

The cumulative default until the end of the loan is equal to the sum of the probabilities of default in each year, as illustrated in Fig. 6.2.

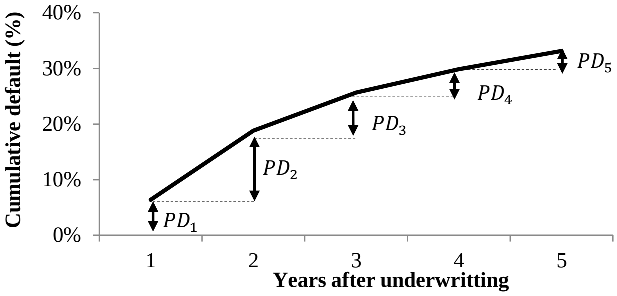


Fig. 6.2: Illustrative example of a cumulative default.

In the remainder of the chapter, we will assume that both the PD and the LGD parameters are available for each borrower at the time of the underwriting.

### 6.2.2. The model

The starting point for achieving the return expected for the loan is the equality:

$$\text{Profits} = \text{Costs} \tag{15}$$

For the costs associated with the loan, we consider the cost of funding and the losses associated with credit default events, the expected loss (EL). Other costs, such as administrative or infrastructure costs, are not considered in this model, despite of being relevant for the banks' activity based costing.

$$\text{Costs}{=}C_E + C_D + EL \tag{16}$$

Profits include the fees and the interests charged on the loan. We assume that the interest rate consists of a market reference rate, $\mathrm{r_r}$, plus a spread, $sp$. The spread component is intended to cover a profit margin and the risks associated with the loan, such as credit risk, liquidity risk, market risk, operational risk, and other risks as may be identified by the bank from time to time. Traditionally, fees, $f$, can be either ad *valorem* or bullet. The former is equivalent to using a spread. The latter can be converted into a spread prior to the calculation of profits, based on the maturity of the loan and the frequency of payments.

$$\text{Profits}{=}(r_r + sp)A + f \tag{17}$$

Considering the equalities in (16) and (17), equation (15) is rewritten as:

$$(r_r + sp)A + f = C_E + C_D + EL \tag{18}$$

equivalent to:

$$(r_r + sp)A + f \;=\; r_E E + r_D D + EL \tag{19}$$

Isolating the components of spread intended to cover the credit and liquidity risks, applied to the total amount to be financed, the previous equation is rewritten as:

$$(r_r + sp_c + sp_l + sp_o)A + f = \; r_E E + r_D D + EL \tag{20}$$

where, $sp_c$ is the credit risk spread, $sp_l$ is the liquidity spread, and $sp_o$ is the spread that covers the profit margin and remaining risks, like operational or country risk.

### 6.2.2.1. Credit risk spread

Equation (20) can be rewritten as:

$$sp_c A = \; r_E E + r_D(A - E) + EL - (r_r + sp_l + sp_o)A - f \tag{21}$$

Rearranging the arguments and isolating the credit risk spread, this is equivalent to:

$$sp_c = \; \frac{E}{A}(r_E - r_D) + (r_D - r_r - sp_l - sp_o) + \frac{EL - f}{A} \tag{22}$$

Considering the equality in (10), then the credit spread that allows an adequate return on equity[6] is:

$$sp_c = sr \times rw(r_E - r_D) + (r_D - r_r - sp_l - sp_o) + \frac{EL - f}{A} \tag{23}$$

If the bank funds the debt by the market reference rate, i.e. $r_D = r_r$, then the credit risk spread is:

$$sp_c = sr.rw.(r_E - r_r) - (sp_l + sp_o) + \frac{EL - f}{A} \tag{24}$$

### 6.2.2.2. Return on equity of a loan

Rearranging the arguments in equation (20), and isolating $r_E$, then the return on equity rate needed to fund the loan is:

$$r_E = \tfrac{1}{E}(sp_c A - EL) - \tfrac{A}{E}(r_D - r_r - sp_l - sp_o) + r_D + \tfrac{f}{E} \tag{25}$$

Or, stated equivalently:

$$r_E = \frac{1}{sr.rw}\left[(sp_c - r_D + r_r + sp_l + sp_o) - \frac{EL - f}{A}\right] + r_D \tag{26}$$

Likewise, if the bank funds the debt by the market reference rate, i.e. $r_D = r_r$, then the return needed to fund a loan is

$$r_E = \tfrac{1}{sr.rw}\left[(sp_c + sp_l + sp_o) - \tfrac{EL-f}{A}\right] + r_r. \tag{27}$$

[6] Here we consider the return before taxes.

## 6.3. Empirical study

### 6.3.1. Data

The empirical study summarized in this chapter is based on the Freddie Mac's single family mortgage loan-level dataset, first published in March 2013. We have followed the performance of 16.7 million of fully amortized 30-year fixed-rate mortgages loans in the U.S., granted between January 1st 1999 and March 31st 2013. Loans performance has been measured in a monthly basis and, at the time of this research, data for performing loans and those that were up to 180 days delinquent were available through September 30th 2013. The dataset is updated over time, typically at the end of each quarter. Release changes, as well as a general user guide describing the file layout and data dictionary,  are recorded online (Freddie Mac, June 2014a). Freddie Mac's information regarding the key loan attributes and performance metrics can be linked to this research in the aggregated summary statistics (Freddie Mac, June 2014b).

Data of the original datasets were aggregated by the origination year. Scores may vary in the range 300-850, or be unknown. Situations where the score is unknown are described by Freddie Mac's Corporation (June 2013b), who refers that these may be an outcome of the sellers' reduced level of verification. We divided the range of possible scores into equidistant intervals of 25, except for the lower and upper bounds. To have dimension, these bounds were aggregated in the buckets [300, 550[ and [800, 850[, respectively.

### 6.3.2. Interest rates setting

Mortgage interest rates are affected by many factors. In the United States they are heavily influenced by the monetary policies of the Federal Reserve Board's Federal Open Market Committee (FOMC). Trends in interest rates on longer financial instruments, such as mortgages, typically follow the fluctuation of the 10-year Treasury note yield. Fig. 6.3 shows the evolution of the 10-year Treasury constant maturity rate over the period 2005 through 2012, and the 10-year constant maturity advance rates authorized by the Federal Housing Finance Agency (FHFA) for the advance pricing.

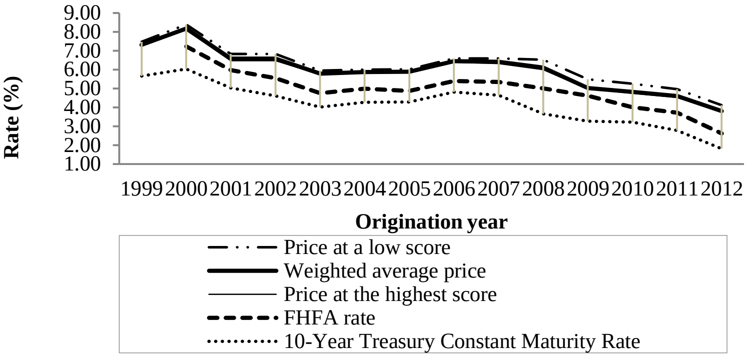


Fig. 6.3: Case study U.S. - Rates evolution in the period 1999-2012.

Sources: Freddie Mac's single-family loan level dataset, FHFA historical advance rates and FED 10-year Treasury note yield.

These rates serve as a reference for the 30-year fixed-rate mortgage loans. Their evolution is represented in Fig. 6.3, together with the average rate, the prime rate and the subprime rate. Borrowers' scores are used to differentiate the interest rates of the mortgages, with the subprime rates being superior to the prime rates. In the period under analysis, the average rates are shifted to the prime rates (in Fig. 6.3. these lines are overlapped), because lending tends to be more driven to borrowers with the highest scores. The gap between the subprime and the prime rates slightly increased in the aftermath of the crisis, between 2008 and 2011, as illustrated in Fig. 6.4.

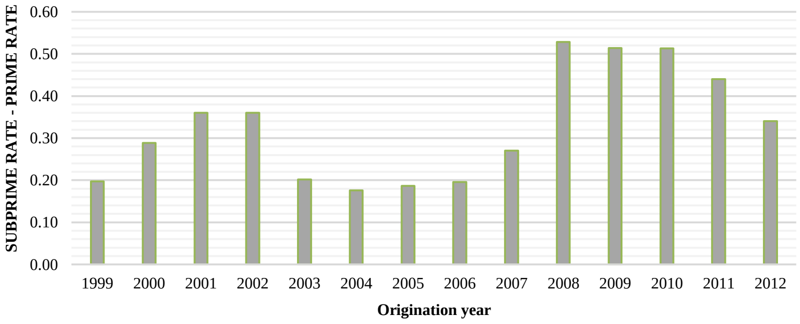


Fig. 6.4: Case study U.S. - Gap subprime - prime rate, 1999-2012.

Source: Freddie Mac single-family loan level dataset.

Between 2004 and 2007, spreads placed around 170 bps for the prime loans and in the average portfolio, and around 190 bps for the subprime loans (Fig. 6.5). Spreads increased in the aftermath of the crisis, and reached a peak in 2008 (Fig. 6.5). A recent report of the Federal Housing Finance Agency says that a 170 bps spread is expected from 2015 through 2017 (Schultz, August 2014) to cover liquidity and credit risks.

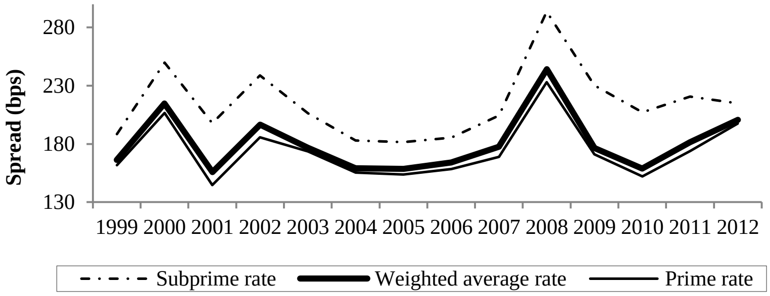


Fig. 6.5: Case study U.S. - 30-year mortgage spreads in the period 1999-2012.

### 6.3.3. Stress testing scenarios

#### 6.3.3.1. Expect margin and expected losses assumptions

We have analysed the performance of the returns in each year after the loan has been underwritten. Losses have been empirically assessed based on the past observed default rates in each score bucket. Rather than just using the average default in the period, as traditional approaches often do, we have focused on the two possible extreme circumstances – an optimistic scenario and a pessimistic scenario. The optimistic scenario is ruled by the cumulative default rates by vintage of an origination year with a lower 1-year default rate (1999). The pessimistic scenario is ruled by the cumulative default rates by vintage of the origination year with the highest 1-year default rate (2008). We have also analysed an average scenario, as a reference, which is ruled on the weighted average annual cumulative default rates in the period 1999-2008. Default rates measured after 2008 were not considered in this analysis because the vintage curves could not be measured from the fifth year onwards (e.g. for the loans underwritten in 2009, only a 4-years vintage can be measured until 2013, and for the loans underwritten in 2012, only a 1-year vintage can be measured until 2013). For the purpose of this study, a borrower is considered to have entered in default after completing 90 consecutive days in delinquency. The vintage curves presented in a previous study of Landy, Ashworth and Yang (2014) suggest that the default rate by vintage have a plateau from the fifth year onwards. Hence, for this analysis we assume the measures of default rates in the first five years of the loan will reveal most of the expected default of the portfolio. The real default rates were measured from the single-family mortgages loan level dataset in each year between 1999 and 2008.

For the three scenarios of default evolution under hypothesis - the average, the optimistic and the pessimistic - we study the performance of the returns along time based on the three corresponding representations of expected losses. Hence, we analyse the performance of the returns under nine potential circumstances, thoroughly illustrated in Fig. 6.6. For simplicity, only the cumulative default curves in the average portfolio and in two distant score buckets are represented – the low score is in the range [600; 625[, and high score is in the range [800; 850[. The three representations of expected losses are:

- *Conjectural model*: assumes that the proportion of new defaults in each year remains constant along the expected life of the asset. In other words, if PD is the probability of default one year after the loan has been granted then the probability of new defaults in the year i is $(1 - PD)^{i-1}PD$.
- *Semi-conjectural model*: The evolution of new defaults is given by the conjectural model until the fifth year, and the model assumes that from that point onwards there are no new defaults.
- *Observed model*: assumes the observed proportion of new default until the fifth year of the loan and that from that point onwards there are no new defaults.

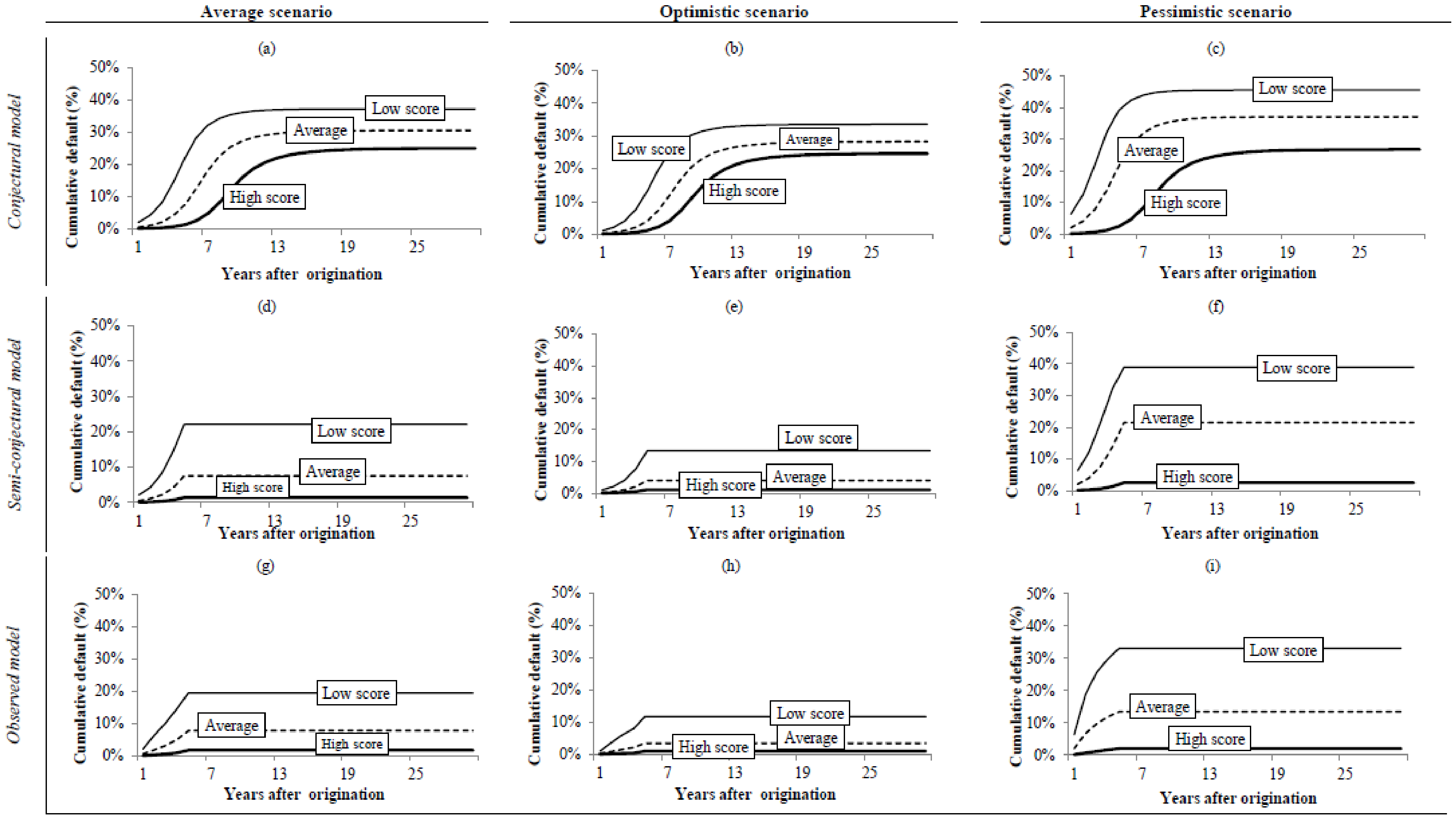


Fig. 6.6. Case study U.S. - Cumulative default rates used in each scenario.

Three scenarios of expected losses were used - average, optimistic and pessimistic – and using three models to represent the cumulative default rate evolution – conjectural, semi-conjectural and observed. The low score is in the range [600; 625[, and high score is in the range [800; 850[. The average is the weighted average cumulative default rate of the portfolio.

We measure the expect returns along time, after the loan has been originated, using the methodology described in section 6.2.1. The calculation of the expected losses has assumed the PD implied in each of the nine scenarios under hypothesis (Fig. 6.6) and retaining the two LDG values behind the Basel Committee's illustrative IRB risk weights[1], 25% and 45% published by the Bank for International Settlements (2006). To calculate the expected margin we considered the average advance rates in the origination year, which are published online by the Federal Home Loan Bank (Federal Home Loan Bank, 2014).

#### 6.3.3.2. Risk-weights parameter assumptions

Under the Basel III capital requirements, countries may define specific rules for the risk-weight parameters.

Many banks using the IRB in the European Union (EU) calculate the risk weights for mortgage loans based on the formula provided by the Basel Committee for the residential retail risk category. And so, the risk weight attached to these mortgages largely depends on the lender's historical default losses experience, subject to cyclical phases' (e.g. downturn) assumptions, which drives the internal risk models. For realistic values of PD and LDG this can give rise to risk weights on this risk category well below 35%.

In the U.S. the Final Rule approved on July 2nd 2013, by the Board of Governors of the Federal Reserve System brought the U.S. banks into

[1] Annex 5, page 279.

compliance with the Basel III capital framework agreed in December 2010. For the banks using the advanced approach in computing risk-based regulatory capital, the final rule took effect in January 1st 2014. For the majority of the U.S. banks operating under the standardized approach, the Final Rule took effect on January 1st 2015. The Final Rule established the standardized approach and the advanced approach to calculate risk-weighted assets in the U.S. However, for Advanced Banks, the standardized approach is considered to establish the minimum generally applicable capital floor requirements for purposes of the section 171 of Dodd-Frank - the Collins Amendment. The Final Rule retained the 50% risk-weight for the residential mortgage loans secured by a first lien on a one-to-four family residential owner-occupied or rented property. This rule does not apply to the loans imprudently underwritten, that are 90 days or more past due or in accrual status, or modified or restructured loans, other than pursuant to the Home Affordable Modification Program.

In this research we aimed at representing the most relevant risk-weights parameters. Hence, to assess the expect returns, we considered two risk-weighting approaches:

- A fixed 50% risk-weight, as used within the U.S. banks for the residential mortgage loans secured by a first lien on a one-to-four family residential owner-occupied or rented property.
- The risk-weights parameters computed with the Basel Committee formula for the residential retail risk category, $12.5 \times 1.06 \times \mathrm{K}$, used within many Advanced banks in the European Union.

## 6.4. Results

Returns are shown for each hypothesized scenario and applying the risk-based model in equation (27). This section we exhibit a set of figures illustrating the returns along the life of the assets, since the loans have been originated until maturity. In each graph, the bottom curves represent the returns in the lower score buckets, beginning in the score 300. Returns' curves move progressively along the score buckets in the direction of the upper curve, which represents the returns in the higher scores bucket [800; 850[. The returns of the aggregated portfolio of loans underwritten in the same year are represented in the bold line in each graph – the average return. Any random pooling of loans of the aggregated portfolio, which have been originated in the same year, should produce the average return. Any portfolio of loans, originated in the same year, in selected ranges of scores should produce a weighted average return of those score ranges. Some combinations of loans with different origination years may produce negative returns. This may occur when the resulting portfolio has a significant proportion of loans that did not reach the break-even point or if it has a significant proportion of loans in specific score ranges that coexist with negative returns in a given point in time.

Our study shows the extent to which return on equity to finance loans largely depends on the rule for calculating the risk-weighted assets. When banks use a constant risk-weight factor, like in the U.S., returns are upper-bounded (meaning that returns cannot be higher than a certain value). Banks calculating the risk-weights under the IRB as defined by the Basel Committee, and operating in markets where the competitors use the standard approach, may have little incentive to decrease prices to customers with the highest scores, by placing their prices in the market prices. If this happens, these banks will certainly reach higher return rates when lending to these segments.

It is now recognised that credit losses arising from credit defaults were far superior than anticipated during the subprime crisis. This was an outcome of an unprecedented decline in home prices that led to a devaluation housing-related securities and rise in foreclosures, together with a sudden escalation of the delinquency rates. The estimation of the LGD parameter depends mostly on the banks' ability in recovering and collecting defaulted loans. Banks in the advanced economies have their own LGD estimates. Banks may use assessment methods ranging from simple LGD estimates at an aggregated level to more sophisticated models, as under the IRB. This information is usually not public, therefore, in this study we have analysed the returns under two standard LGD values of the Basel Committee - 25% and 45% (Bank for International Settlements, 2006). To the extent of our knowledge, for the residential retail exposures, the LGD should be closer to the 25% reference value in average conditions. But this value is expected to rise in times of recession, when banks are swamped with defaulted loans and high provisions. Hence, the 45% LGD reference should better allow replicating the conditions of a catastrophic situation.

For values of LGD up to 25%, most of the credit score buckets produce positive returns along the entire life of the asset, either in average conditions or in adverse circumstances. Few exceptions are remarked in the very lowest scores ranges, the subprime loans, between score 300 and score 625, where the returns turn negative somewhere between the second and the fourth year after the loan has been originated, if the loan is originated under the most adverse scenario. These findings are confirmed in the graphs of the pessimist scenario in Fig 6.7 and Fig 6.8, and more markedly in the observed default model (graphs (i)). However, returns turn positive from the fifth year onwards either if we assume that the cumulative defaults reach a plateau by the fifth year (Fig 6.7 (f), Fig 6.7 (i), Fig 6.8 (f) and Fig 6.8 (i)) or if we consider that the new default evolves constantly over time until loan maturity (Fig 6.7 (c) and Fig 6.8 (c)). A portfolio of loans randomly selected from the aggregated portfolio, originated in the same year, generates positive returns along the entire life of the assets (see the bold curve is the graphs of all scenarios in Fig 6.7 and Fig 6.8).

Placing the LGD in 45% suggests a discussion from a very different perspective. A higher LGD amplifies the disturbances stimulated by the increase in default rate, and so, under this setting, a significant number of loan pooling arrangements can generate negative returns. This is valid throughout the cycle, in average conditions, but is more applicable to extreme adverse circumstances, because an LGD of 45% is more likely to occur during times of serious financial distress. Hence, we centre our discussion in the results for the pessimistic scenario. In this setting, a portfolio of loans randomly selected, originated in the same year, should produce a weighted average return, represented in the bold curves in Fig 6.9 and Fig 6.10, which even under the pessimistic scenario are able to produce positive returns. However, in the years of sharp increase in default rates, returns are negative, which may occur somewhere between the third (Fig 6.9 (i) and Fig 6.10 (i)) and the fifth year (Fig 6.9 (c), Fig 6.9 (f), Fig 6.9 (c) and Fig 6.10 (f)). When isolating the returns in each score bucket, a wide range of scores has negative returns in the first years of the loan. In particular, up to the score 675, returns are below the average. In the score range [300; 575[ returns are negative between the origination point and the fifth year. In the score range [575; 675[ returns are negative between the first and sixth year, depending in the scenario of default evolution (conjectural, semi-conjectural or observed). When the new defaults evolve as defined under the conjectural default model, i.e. assuming that the new defaults evolve constantly over time, then the best scored loans also reach negative returns is some point in time, until the tenth year of the loans (Fig 6.9 (c) and Fig 6.10 (c)). This is more noticeable when using the risk-weights calculated with the Basel Committee formula (Fig 6.10 (c)). Under this setting, the loans with the highest scored borrowers may produce highly negative returns around the tenth year.

Following the subprime crisis, lenders are firmly declining the subprime loans, below the score 620, as described in a paper of Keys et al (2008). In fact, the risk of these loans is higher than in other scores, as previously stated. However, when these loans go older, mostly after the fifth year, the expected returns increase.

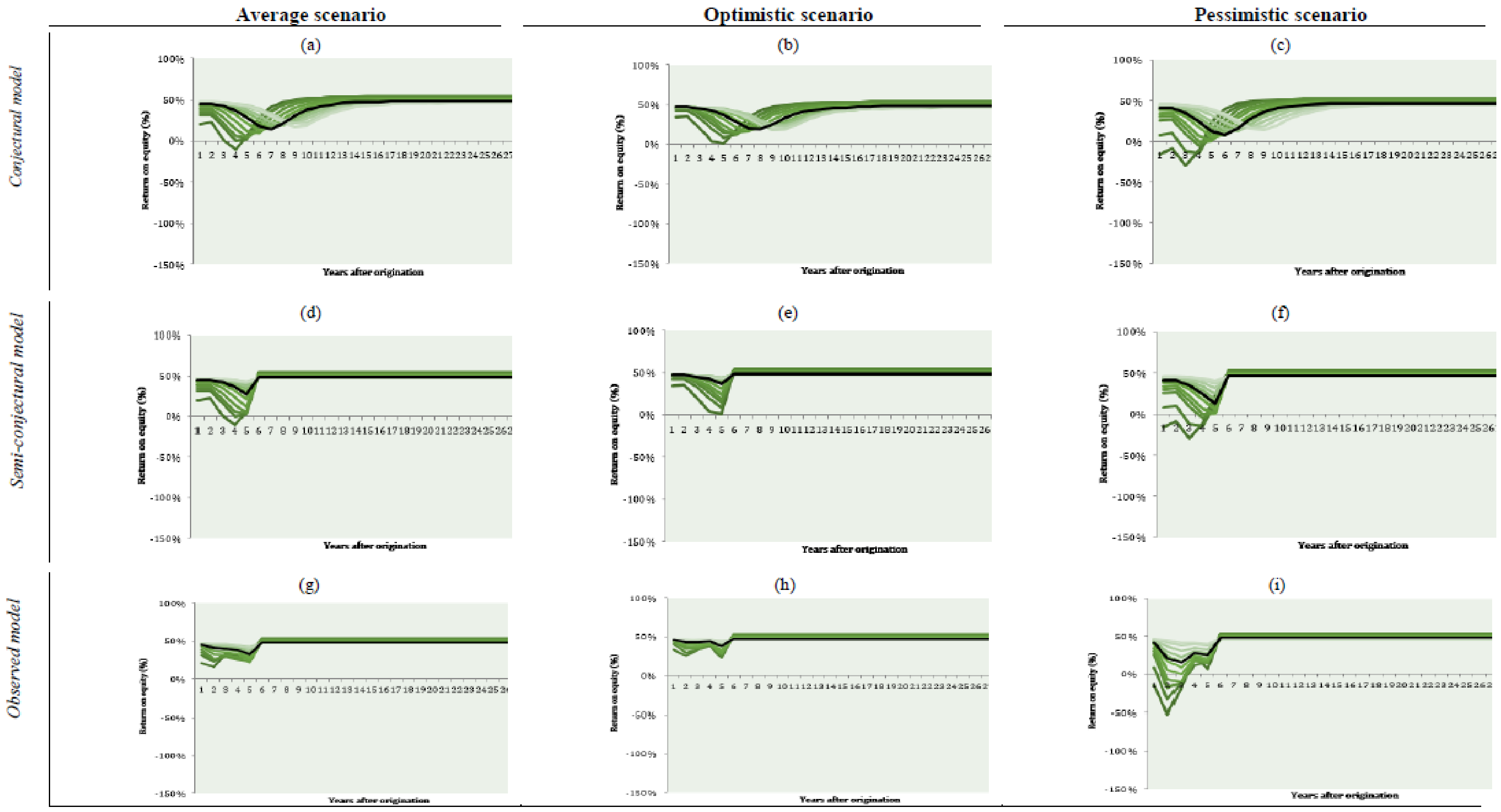


Fig. 6.7. Case study U.S. - Expected returns, rw=50% and LGD=25%.
From down to the top of the image, lines represent the expected returns in the score buckets [300; 550[, [550; 575[,[575; 600[, [600; 625[,[625; 650[, [650; 675[,[675; 700[, [700; 725[, [725; 750[, [750; 775[,[775; 800[, [800; 850[, respectively.

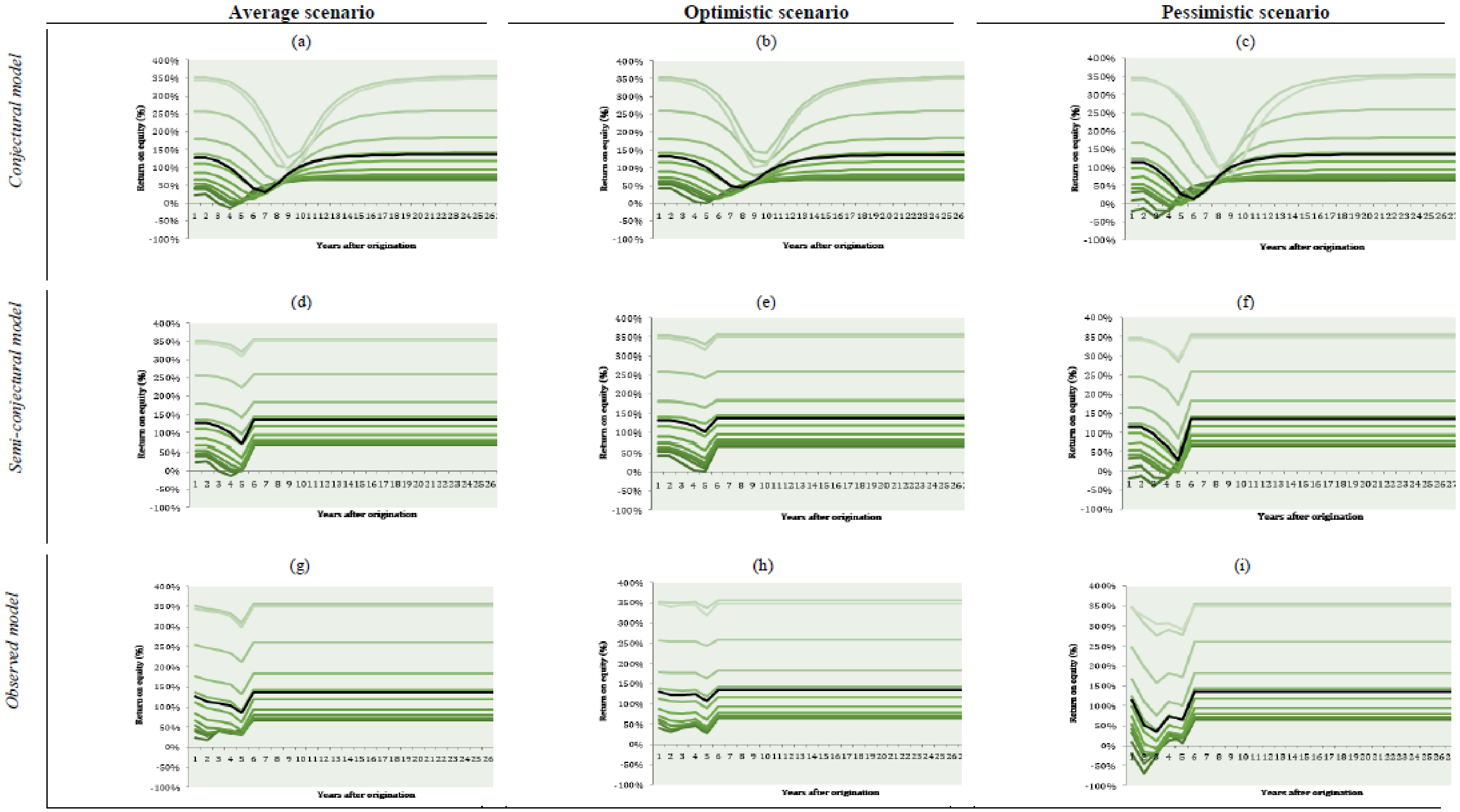


Fig. 6.8. Case study U.S. - Expected returns, rw=12.5×1.06×K and LGD=25%.
From down to the top of the image, lines represent the expected returns in the score buckets [300; 550[, [550; 575[,[575; 600[, [600; 625[,[625; 650[, [650; 675[,[675; 700[, [700; 725[, [725; 750[, [750; 775[,[775; 800[, [800; 850[, respectively.

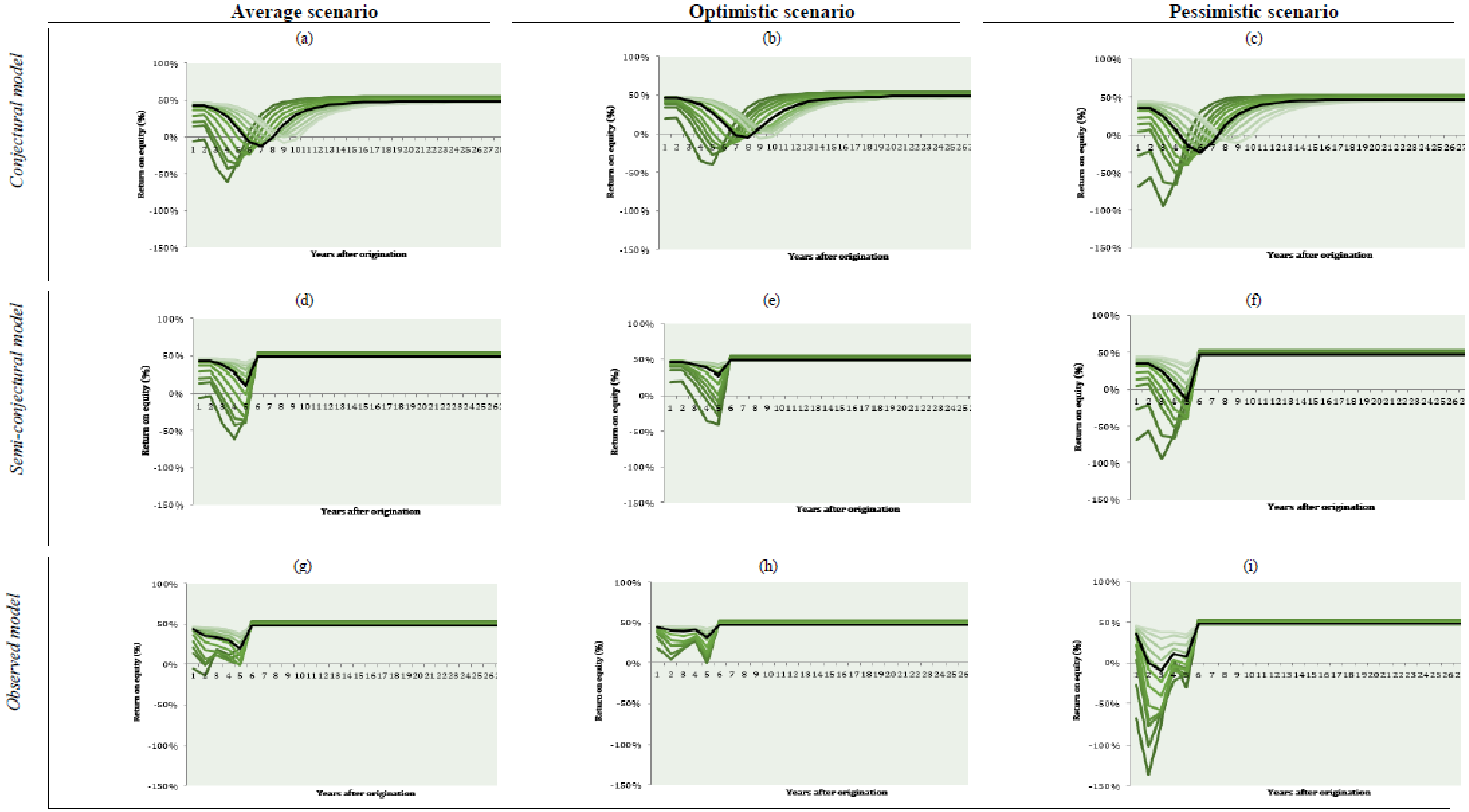


Fig. 6.9. Case study U.S. - Expected returns, rw=50% and LGD=45%.
From down to the top of the image, lines represent the expected returns in the score buckets [300; 550[, [550; 575[,[575; 600[, [600; 625[,[625; 650[, [650; 675[,[675; 700[, [700; 725[, [725; 750[, [750; 775[,[775; 800[, [800; 850[, respectively.

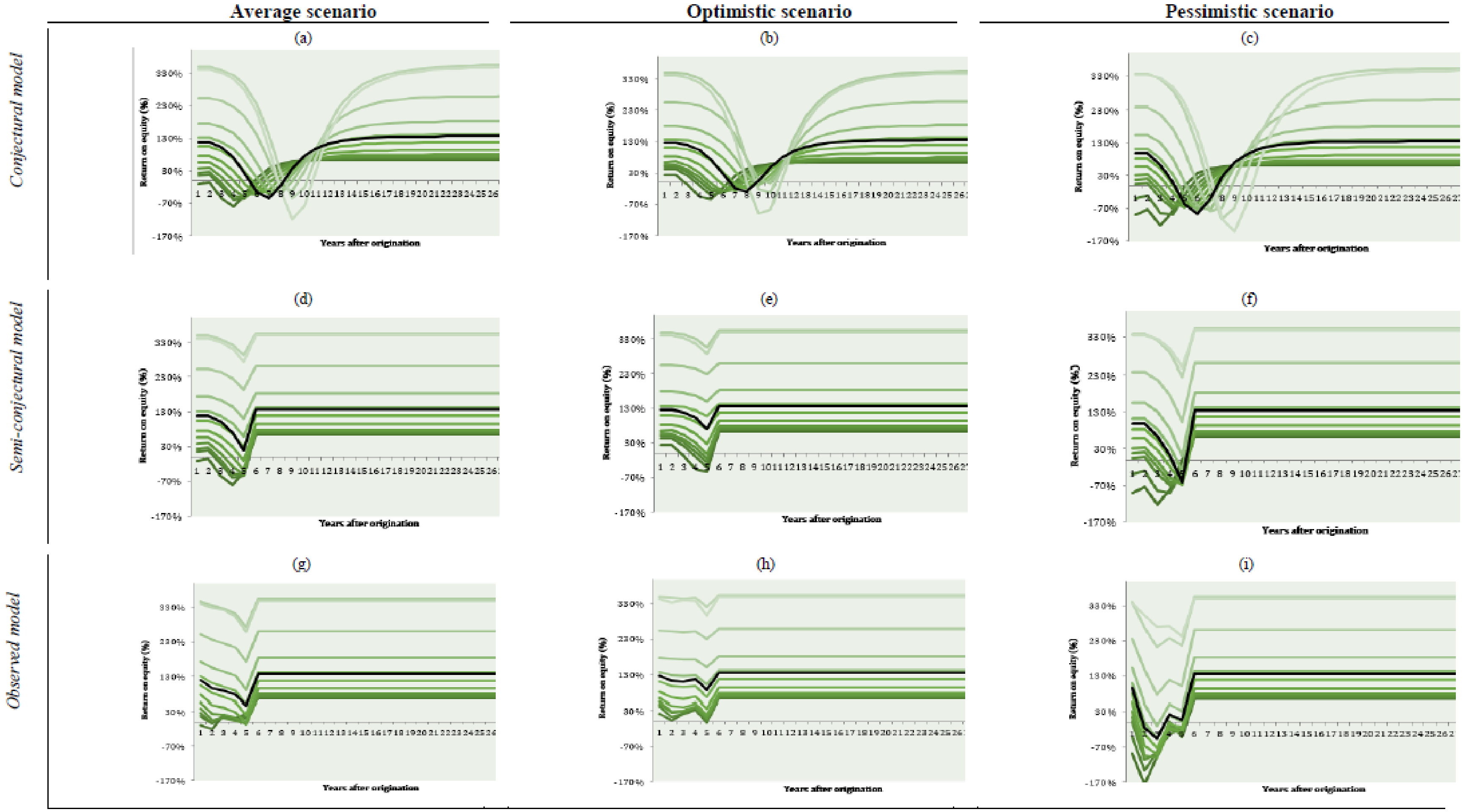


Fig. 6.10. Case study U.S. - Expected returns, rw=12.5×1.06×K and LGD=45%.
From down to the top of the image, lines represent the expected returns in the score buckets [300; 550[, [550; 575[,[575; 600[, [600; 625[,[625; 650[, [650; 675[,[675; 700[, [700; 725[, [725; 750[, [750; 775[,[775; 800[, [800; 850[, respectively.

## 6.5. Conclusions

Regulators, academics and the financial industry increasingly recognize the importance of stressing their reference models under extreme circumstances. Stress-testing often rely on theoretical assumptions or heavily depend on projections (e.g. GDP growth, unemployment or the business in the following period), which in turn are based in several other conjectures. This approach may be of little value if the premises are inaccurate or if the projections remain valid during a short-term period. The former may lead to misleading conclusions; the later may turn outdated fairly quickly. If, on the one hand, the practical value of these exercises may be debatable, on the other, understanding how real systems evolve in normal conditions and during exacerbated circumstances is of a great use.

We present a stress-testing exercise based in a real-world dataset of 16.7 million loans that were at the epicentre of the global crisis, the Freddie Mac's single family mortgage loan-level dataset, first published in March 2013. We did not attempt to draw a set of baseline assumptions for the future macroeconomic or business conditions. Instead, we simulated the most extreme circumstances of the past, which impacted severely in credit risk and in returns. The measure of risk uses the FICO score, which is an industry standard in the U.S., currently used in 90% of the lending decisions to determine how much money each individual can borrow and to set the interest rate for each loan. Hence, since its introduction 20 years ago, the latest financial crisis might have been the most severe disruption affecting the borrower's ability to repay their debt.

We analysed the performance of the returns over time considering nine potential scenarios of default rate evolution, each was replicated for an LGD of 25% and LGD of 45%, and for a constant 50% risk-weight and applying the risk-weight formula of the Basel Committee for the IRB approach. With very few exceptions, for an LGD up to 25%, most of the credit score buckets produce positive returns along the entire life of the asset, either in average conditions or in adverse circumstances. Within this setting, a portfolio of loans randomly selected from the aggregate portfolio, originated in the same year, should generate positive returns. Higher LGD amplifies the disturbances caused by default rate increases. For a 45% LGD, a significant number of loan pooling arrangements can generate negative returns, either through the cycle, but more noticeably under extreme adverse circumstances. Under this setting, for a wide range of scores, up to 675, returns are below the average in the first years of the loans. If the new default rates evolve constantly over time, then the best scored loans may also reach negative returns is some point in time, in the first 10 years of the loans.

Following the subprime crisis, lenders are firmly declining the subprime loans, below the score 620, and restricting credit to the adjacent lower score borrowers. Although the risk of these loans is higher than in other scores, more consideration should be promoted around the expected return of these loans, including the subprime. When these loans go older, mostly after the fifth year, the expected return increase, meaning that in the long run, the loans in the lowest scores can positively contribute to the overall return of the portfolios. Although we did not analyse the effect of early repayments, intuition and practical knowledge also suggest that these loans will probably not be paid in advance, and so they are valuable for the portfolios' compositions. So, current regulation would be improved if specific rules were developed for accurately pricing loans to the low scores borrowers, rather than strictly prohibiting. This could be complemented by imposing boundaries in the proportion of these loans in the composition of the portfolios, bearing in mind that the risk significantly reduces when these loans go older, in contradiction to the highest score loans that may reach peaks of negative returns around the tenth year. In so doing, the target market, which now is concentrated in the prime scores borrowers, would expand, which is crucial in retail banking.

# 7. Conclusions

## 7.1. Summary and contributions

The main contribution of this research is a new modelling framework for credit risk assessment that extends the prevailing credit scoring models built upon historical data static settings. We also extended a line of investigation consisting of a stress-testing methodology that was applied to the Freddie Mac's database.

In the core of our research, we propose an initial approach to dynamic credit scoring in which the credit scoring model is developed with a static modelling framework, and then adjusted by time-changing macroeconomic factors. Afterwards, we extended the previous methodology by proposing a new dynamic modelling framework based on a sequential learning of the new incoming data (Sousa et al., 2016). Within the new modelling schema, predictions are made upon the sequence of temporal data, which is more suitable for adapting to real default concept drifts, translated by changes in the population, in the economy or in the market. The new framework enabled us to depict the two basic mechanisms of memory - short-term (STM) and long-term memory (LTM) - in credit risk assessment, with a set of empirical studies using two real-world financial datasets.

The first empirical studies used a sample of credit cards of a financial institution operating in Brazil from 2009 through 2011. Then, we enhanced the research with the Freddie Mac's database, available since 2013, where we analysed 16.7 million loans in the U.S. that were at the epicentre of the global crisis, granted between 1999 and 2013. Three plain assumptions are confirmed in our investigation: newest data consistently improves forecasting accuracy; STM allows a quick adaptation to changes and LTM is more favourable in stable conditions. In real-world environments, different amounts of memory can be explored concurrently. This is important in the credit risk area, which often undergoes shocks. During a shock, limited memory is important; other times a larger memory is favoured.

In the banking industry, credit scoring models are developed from static windows and often kept unchanged over years. In this setting, the two basic mechanisms of memory, STM and LTM, which are fundamental to learning, are still overlooked in current modelling frameworks. As a consequence, they are insensitive to changes, like population drifts or financial distress. The usual outcomes are the default rates rising and abrupt credit cuts, as those that were observed in the U.S., in the aftermath of the last Global crisis (Sousa et al., 2015a). This is one area where new thoughts, like simplifying current decision layers, need to be encouraged, because regulators still promote models whose coefficients do not change over time.

Finally, being aware that no rating system is fully capable of anticipating distressing events, we developed a research pipeline, consisting of a stress-testing methodology that was applied to the Freddie Mac's database. We study how the return on lending evolves in normal conditions and under extreme adverse circumstances, subject to the current capital regulations, under the Basel Accords. As a part of our major contributions, we settled a first return on a risk-adjusted equity model embedded in the contemporary capital regulatory framework. Then we proposed a stress-testing methodology where the driving factors are the default by vintage and LGD, an alternative to using heavy and uncertain theoretical or macroeconomic assumptions. We present a first application of Freddie Mac's data to stress-testing, which can easily be replicated in other real-world consumer finance portfolios. This allowed us to explicitly exhibit the impact of PD and LGD in the return on lending, based on the Freddie Mac's mortgage loans, and to argue that the sudden credit-cuts by score have been an overreaction to the last global crisis. Under certain values of LGD, empirical simulations show that lending to borrowers with lower scores produces positive returns in the long-run. If sufficiently mature, these loans can boost portfolios compositions because they are less exposed to early payments. We claim that regulators and lenders should ascertain the LGD boundaries under which the bank operates to drive lending policies in retail banking.

## 7.2. Limitations and research directions

In the main line of investigation, there are some important topics in the default concept drift that we did not consider, which we have deferred for future research. Firstly, some specifics of the financial contexts may seem fairly stable along some periods. While this thesis provides convincing results, some additional simulations using real-world datasets and longer time frames would be valuable. Secondly, modelling the delinquency presents challenges, since a window of time is required in order to measure the outcome, that is, the true class, before the new model is built. Therefore, for forecasting, it follows that there will be a time gap of the same length between the values of predictor variables used in the model and the first possible forecast period in the future. Although this is not a problem of the proposed methodology, future research should bring new insights to overcome this issue, with a view on practicality. Thirdly, some good alternatives to using windows of data blocks are encouraged, which may be based on using ensembles of the models learnt in the past, possible combining the two components of memory, short-term and long-term memory, or a forgetting factor method. In relation to the STM component, a prior selection of the window length seems appropriate and could be employed to optimize the adaptation ability. There is some material on this going back to Adams (2010). Fourthly, our empirical study considered a set of fixed predictors. Future research should consider sets of predictor of variable length. This is important for detecting concept drift because the set of predictors that are being used may be quite limited to exhibit signs of change, even if they are occurring in the environment. Finally, performance is reported in this paper, but the conditions leading to differences in performance are not fully explored.

In the pipeline research, we did not address some important subjects, into which an in-depth investigation is encouraged.

Firstly, the model ignores the prepayment and refinancing risks. Although our expectation is that, in doing so, our findings would be strengthened because early repayments should be shifted to higher scored borrowers. Some good ideas on how to introduce these risks in the model could be promoted and their impact in the results analysed. Some key concepts relevant for this task are proposed in a recent paper of Campbell and Cocco (2015).

Secondly a key parameter in the comparative analysis is the loss given default (LGD). Our simulations use an LGD which is uniformly distributed across credit risk classes or, in other words, the LGD and the credit score are independent variables. Therefore, some additional modelling in the LGD parameter could be valuable.

Finally, we consider the credit risk and the cost of funding evaluations at the time of the application: a period which may be too short considering the maturity of the credit. So, time-varying evaluations would improve the realism of these simulations. As far as we could ascertain, there is little research on how to solve this practical limitation. We believe that the methodology developed in our main line of investigation can be extended to address this problem. This is another direction for future research.

# References

ADAMS, N. M., TASOULIS, D. K., ANAGNOSTOPOULOS, C. & HAND, D. J. 2010. Temporally-Adaptive Linear Classification for Handling Population Drift in Credit Scoring. *Proceedings of COMPSTAT'2010*.

ALTMAN, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23**,** 589-609.

AMATO, J. D. & FURFINE, C. H. 2004. Are credit ratings procyclical? *Journal of Banking & Finance*, 28**,** 2641-2677.

ANDERSON, R. 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press.

ANDERSON, S. & JOZWIK, J. 2014. Building a Credit Model Using GSE Loan-Level Data. *The Journal of Structured Finance*, 20**,** 19-36.

ANTÃO, P. & LACERDA, A. 2011. Capital requirements under the credit risk-based framework. *Journal of Banking & Finance*, 35**,** 1380-1390.

AVERY, R. B., CALEM, P. S. & CANNER, G. B. 2004. Consumer credit scoring: do situational circumstances matter? *Journal of Banking & Finance*, 28**,** 835-856.

BADDELEY, A. 2012. Working memory: theories, models, and controversies. *Annual review of psychology*, 63**,** 1-29.

BANCO CENTRAL DO BRASIL 2011. Relatório sobre a Indústria de Cartões de Pagamentos Adendo Estatístico. *In:* SECRETARIA DE ACOMPANHAMENTO ECONÔMICO – MINISTÉRIO DA FAZENDA, S. D. D.

E. M. D. J. (ed.). Departamento de Operações Bancárias e de Sistema de Pagamentos.

BANK FOR INTERNATIONAL SETTLEMENTS 2004. Implementation of Basel II: Practical Considerations. *Basel Committee on Banking Supervision, Basel*.

BANK FOR INTERNATIONAL SETTLEMENTS 2006. International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version. *Basel Committee on Banking Supervision, Basel*.

BANK FOR INTERNATIONAL SETTLEMENTS 2010. Basel III: A global regulatory framework for more resilient banks and banking systems. *Basel Committee on Banking Supervision, Basel*.

BEHN, M., HASELMANN, R. & WACHTEL, P. 2015. Procyclical Capital Regulation and Lending. *The Journal of Finance*.

BELLOTTI, T. & CROOK, J. 2009. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society,* 60**,** 1699-1707.

BIANCO, K. M. 2008. The subprime lending crisis: Causes and effects of the mortgage meltdown.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. & STONE, C. 1984. Classification and regression trees. *Wadsworth International Group: Belmont, California*.

CAMPBELL, J. Y. & COCCO, J. F. 2015. A model of mortgage default. *The Journal of Finance,* 70**,** 1495-1554.

CHEN, M.-C. & HUANG, S.-H. 2003. Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications,* 24**,** 433-441.

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

CROOK, J. N., EDELMAN, D. B. & THOMAS, L. C. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research,* 183**,** 1447-1465.

CROOK, J. N., THOMAS, L. C. & HAMILTON, R. 1992. The degradation of the scorecard over the business cycle. *IMA Journal of Management Mathematics,* 4**,** 111-123.

DESAI, V. S., CROOK, J. N. & OVERSTREET JR, G. A. 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research,* 95**,** 24-37.

DURAND, D. 1941. *Risk elements in consumer instalment financing,* National Bureau of Economic Research, Inc.

EBA 2013a. Interim results of the EBA review of the consistency of risk-weighted assets. Top-down assessment of the banking book. European Banking Authority.

EBA 2013b. Report on the comparability of supervisory rules and practices. European Banking Authority.

EINAV, L., JENKINS, M. & LEVIN, J. 2013. The impact of credit scoring on consumer lending. *The RAND Journal of Economics,* 44**,** 249-274.

EISENBEIS, R. A. 1978. Problems in applying discriminant analysis in credit scoring models. *Journal of Banking & Finance,* 2**,** 205-219.

FAWCETT, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** 861-874.

FEDERAL HOME LOAN BANK 2014. Fixed-Rate Advances. *In:* BANK, F. H. L. (ed.) *http://members.fhlbdm.com/member-tools/advance-rates/3/, accessed in 02-09-2014.* US.

FEDERAL HOUSING FINANCE AGENCY 2013. Conservatorship Strategic Plan: Performance Goals for 2013. Federal Housing Finance Agency.

FICO 2006. A Fair Isaac White Paper - Introduction to Scorecard for FICO Model Builder. *Fair Isaac Corporation.*

FISHER, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics,* 7**,** 179-188.

FREDDIE MAC June 2013a. Single Family Loan-Level Dataset. Freddie Mac.

FREDDIE MAC June 2013b. Single Family Loan-Level Dataset General User Guide Freddie Mac.

FREDDIE MAC June 2014a. Single Family Loan-Level Dataset - Release Notes. Freddie Mac.

FREDDIE MAC June 2014b. Single Family Loan-Level Dataset - Summary Statistics. Freddie Mac.

GAMA, J. 2010. *Knowledge discovery from data streams,* London, Chapman & Hall/CRC.

GAMA, J., ŽLIOBAITĖ, I., BIFET, A., PECHENIZKIY, M. & BOUCHACHIA, A. 2014. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR),* 46**,** 44.

GOODMAN, L. S., LANDY, B., ASHWORTH, R. & YANG, L. 2014. A Look at Freddie Mac’s Loan-Level Credit Performance Data. *The Journal of Structured Finance*, 19**,** 52-61.

HAND, D. J. 2006. Classifier Technology and the Illusion of Progress *Statistical Science,* 21**,** 30-34.

HAND, D. J. & ANAGNOSTOPOULOS, C. 2013. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34**,** 492-495.

HAROLD BIERMAN, J. & HAUSMAN, W. H. 1970. The Credit Granting Decision. *Management Science,* 16**,** B-519-B-532.

HULL, J. C. 2009. The credit crunch of 2007: what went wrong? Why? What lessons can be learned? *Journal of Credit Risk*, 5**,** 3-18.

JENSEN, H. L. 1992. Using neural networks for credit scoring. *Managerial Finance*, 18**,** 15-26.

JONES, S., JOHNSTONE, D. & WILSON, R. 2015. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance*, 56**,** 72-85.

JUNG, K. M., THOMAS, L. C. & SO, M. C. 2015. When to rebuild or when to adjust scorecards. *Journal of the Operational Research Society,* 66**,** 1656-1668.

KELLY, M. G., HAND, D. J. & ADAMS, N. M. 1999. The impact of changing populations on classifier performance. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM.

KEYS, B. J., MUKHERJEE, T. K., SERU, A. & VIG, V. 2008. Did securitization lead to lax screening? *Evidence from Subprime Loans (December 25, 2008). EFA.*

KLINKENBERG, R. 2004. Learning drifting concepts: Example selection vs. example weighting. *Intelligent data analysis,* 8**,** 281-300.

LAZARESCU, M. M., VENKATESH, S. & BUI, H. H. 2004. Using multiple windows to track concept drift. *Intelligent data analysis,* 8**,** 29-59.

LEE, T.-S. & CHEN, I.-F. 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications,* 28**,** 743-752.

LEE, T.-S., CHIU, C.-C., LU, C.-J. & CHEN, I.-F. 2002. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications,* 23**,** 245-254.

LESSMANN, S., BAESENS, B., SEOW, H. & THOMAS, L. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research,* 247**,** 124-136.

LI, S.-T., SHIUE, W. & HUANG, M.-H. 2006. The evaluation of consumer loans using support vector machines. *Expert Systems with Applications,* 30**,** 772-782.

LUCAS, A. 2004. Updating scorecards: removing the mystique. *Readings in Credit Scoring: Foundations, Developments, and Aims. Oxford University Press: New York***,** 93-109.

MALHOTRA, R. & MALHOTRA, D. K. 2002. Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research,* 136**,** 190-211.

MALIK, M. & THOMAS, L. C. 2012. Transition matrix models of consumer credit ratings. *International Journal of Forecasting,* 28**,** 261-272.

MALOOF, M. A. & MICHALSKI, R. S. 2004. Incremental learning with partial instance memory. *Artificial intelligence,* 154**,** 95-126.

MARTENS, D., DE BACKER, M., HAESEN, R., VANTHIENEN, J., SNOECK, M. & BAESENS, B. 2007. Classification with ant colony optimization. *Evolutionary Computation, IEEE Transactions on,* 11**,** 651-665.

MIN, J. H. & LEE, Y.-C. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications,* 28**,** 603-614.

MYERS, J. H. & FORGY, E. W. 1963. The development of numerical credit evaluation systems. *Journal of the American Statistical Association,* 58**,** 799-806.

NELDER, J. A. & WEDDERBURN, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General),* 135**,** 370-384.

ORTH, W. 2013. Multi-period credit default prediction with time-varying covariates. *Journal of Empirical Finance,* 21**,** 214-222.

PAVLIDIS, N., TASOULIS, D., ADAMS, N. & HAND, D. 2012. Adaptive consumer credit classification. *Journal of the Operational Research Society,* 63**,** 1645-1654.

QUINLAN, J. R. 1986. Induction of decision trees. *Machine learning,* 1**,** 81-106.

RAJAN, U., SERU, A. & VIG, V. 2015. The failure of models that predict failure: Distance, incentives, and defaults. *Journal of Financial Economics,* 115**,** 237-260.

RÊZÁC, M. & RÊZÁC, F. 2011. How to Measure the Quality of Credit Scoring Models. *Finance a Uver: Czech Journal of Economics & Finance,* 61**,** 486-507.

RUTHENBERG, D. & LANDSKRONER, Y. 2008. Loan pricing under Basel II in an imperfectly competitive banking market. *Journal of Banking & Finance,* 32**,** 2725-2733.

SALGANICOFF, M. 1997. Tolerating concept and sampling shift in lazy learning using prediction error context switching. *Artificial Intelligence Review,* 11**,** 133-155.

SAUNDERS, A. & ALLEN, L. 2010. *Credit risk management in and out of the financial crisis: New approaches to value at risk and other paradigms,* John Wiley & Sons.

SCHLIMMER, J. C. & GRANGER JR, R. H. 1986. Incremental learning from noisy data. *Machine learning,* 1**,** 317-354.

SCHULTZ, J. August 2014. The Size of the Affordable Mortgage Market: 2015-2017 Enterprise Single-Family Housing Goals. *Federal Housing Finance Agency.* FHFA.

SILVA, F. B. S. & CARDOSO, J. S. 2015. Differential Scorecards for Binary and Ordinal data. *Intelligent data analysis,* 19**,** 6, 1391-1408.

SMITH, P. F. 1964. Measuring Risk on Consumer Instalment Credit. *Management Science,* 11**,** 327-340.

SOUSA, M. R., GAMA, J. & BRANDÃO, E. 2013. Introducing time-changing economics into credit scoring. *FEP working paper*. University of Porto, Portugal, School of Economics and Management.

SOUSA, M. R., GAMA, J. & BRANDÃO, E. 2015a. Links between Scores, Real Default and Pricing: Evidence from the Freddie Mac's Loan-level Dataset. *Journal of Economics, Business and Management,* 3**,** 1106-1114.

SOUSA, M. R., GAMA, J. & BRANDÃO, E. 2015b. Stress-testing the return on lending under real extreme adverse circumstances. *European Financial Management Association annual conference.* Amsterdam: EFMA.

SOUSA, M. R., GAMA, J. & BRANDÃO, E. 2016. A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications,* 45**,** 341–351.

STEENACKERS, A. & GOOVAERTS, M. 1989. A credit scoring model for personal loans. *Insurance: Mathematics and Economics,* 8**,** 31-34.

SUN, J. & LI, H. 2011. Dynamic financial distress prediction using instance selection for the disposal of concept drift. *Expert Systems with Applications,* 38**,** 2566-2576.

THOMAS, L. C. 2010. Consumer finance: challenges for operational research. *Journal of the Operational Research Society,* 61**,** 41-52.

THOMAS, L. C., EDELMAN, D. B. & CROOK, J. N. 2002. *Credit Scoring and Its Applications,* Philadelphia, Society for Industrial and Applied Mathematics.

TSYMBAL, A. 2004. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin,* 106.

WEST, D. 2000. Neural network credit scoring models. *Computers & Operations Research,* 27**,** 1131-1152.

WIDMER, G. & KUBAT, M. 1993. Effective learning in dynamic environments by explicit context tracking. *Machine Learning: ECML-93.* Springer.

WIDMER, G. & KUBAT, M. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning,* 23**,** 69-101.

YANG, Y. 2007. Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research,* 183**,** 1521-1536.

ZANDI, M. 1998. *Incorporating economic information into credit risk underwriting,* Chicago and London, Fitzroy Dearborn Publishers.