The roles of short-term and long-term memory in credit scoring: evidence from the Freddie Mac's database*

Maria Rocha Sousa[a,1], João Gama[a,b], Elísio Brandão[a]

[a]*School of Economics and Management, University of Porto*

[b]*LIAAD-INESC TEC*

Abstract

We investigate two mechanisms of memory, short-term (STM) and long-term memory (LTM), in the context of credit risk assessment. In this context, memory refers to the amount of historic data being used for estimation. This is important in the credit risk area, which often seems to undergo shocks. These components are fundamental for learning, but are often not considered in credit scoring modelling frameworks. As a consequence, these models are insensitive to changes, like population drifts or financial distress. We use a dynamic framework, extending the prevailing credit scoring models built upon static settings. A key contribution of this paper is provided by the insight that different amounts of memory can be explored concurrently. During a shock, limited memory is important. Other times, a larger memory has merit. Empirical study relying on the Freddie Mac's database, with 16.7 million mortgage loans originated in the U.S. through 1999 to 2013, suggests using a dynamic modelling of STM and LTM components as a way of optimizing current rating frameworks.

Keywords: Freddie Mac, mortgage loans, FICO score, PD, STM, LTM, credit risk, scoring, drift, adaptive learning, temporal degradation

JEL: C5, C61, C63, C67, C8

1. Introduction

More than half-century has passed since credit scoring models have been introduced in credit risk assessment and in the prediction of corporate bankruptcy (Harold Bierman and Hausman, 1970, Altman, 1968, Smith, 1964, Myers and Forgy, 1963). Nowadays, in the advanced economies, a high proportion of the loan applications 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 (U.S.) the measure of risk FICO score is an industry standard, claimed to be 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. In the OECD countries, banks that have adopted the Internal Ratings Based (IRB) Approach, in Basel II Accord (BCBS, 2006, BIS, 2004), use their own credit scoring models at the basis of the minimum regulatory capital calculation.

A credit scoring model is an intelligent system. The output is a prediction of a given entity, current or potential borrower, entering in default in a future period. A representation that gained sympathy in real-world financial environments is 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, 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, as it can be seen in the paper of Jones et al. (2015). Typical credit scoring systems are developed from static data sets. 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 to historical examples within

[1] Corresponding author
Email addresses: 100427011@fep.up.pt (Maria Rocha Sousa), jgama@fep.up.pt (João Gama), ebrandao@fep.up.pt (Elísio Brandão)

such timeframe, the model is designed using a supervised learning approach. The resulting model is then used, possibly for several years, without further relearning. As a consequence, traditional static credit scoring models are quite insensitive to changes in the financial environments, like gradual or abrupt population drifts 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 excess sensitivity to the business cycle.

To some extent, credit scoring modelling still needs to better mimic the human learning established on the experience. There are two basic mechanisms of memory, short-term memory (STM) and long-term memory (LTM) that are fundamental components of the experience and human cognition. The former is easy to set up but readily forgotten; the later 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 on which type of memory configuration of the learning of credit scoring systems enables a rapid adaptation to changes. Hence, our analysis is set on two research questions. Is the recent information relevant to improve forecasting accuracy? Does older information always improve forecasting accuracy?

New concepts for adapting to changes 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 new incoming data. The driving idea mimics the principle of films, by composing the model with 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 more robust predictions in stable conditions, but also in the presence of changes, while the prevailing static frameworks cannot.

Renewed empirical credit risk measures are presented in this paper 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 the historical 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 if an adaptive learning framework could also be undertaken. In such setting, we provide the insight of using a multicomponent memory approach, consisting on 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). This is important in the credit risk area, which often seems to undergo shocks.

1.1. Previous research

This paper is a large extension of a previous research that delivered the winning model in the BRICS 2013 competition in data mining and finance (Sousa et al., 2013). This competition, opened to academics and practitioners, was focused on the development of a credit risk assessment model to deal with the problem of the performance degradation over time, potentially caused by gradual market changes along a few years of business operation. Participants were encouraged to use any modelling technique, under a temporal degradation or concept drift perspective. The winning solution consisted of a two-stage model for dealing with the temporal degradation of credit scoring models, which produced motivating results in a 1-year horizon. The authors first developed a credit scoring model using a set of supervised learning methods, and then calibrated the output, based on a projection of the evolution in the default. This adjustment considered both the evolution of the default and the evolution of macroeconomic factors, echoing potential changes in the population of the model, in the economy, or in the market. In so doing, adjusted scores combined the customers' specific risk with systemic risk. The winning team concluded that the performance of the models did not significantly differ among classification models, like logistic regression (LR), AdaBoost, and Generalized Additive Models (GAM). However, after training in several windows lengths, they observed that the model based on the longest window has produced the best performing model in the long-run, among all competitors. This finding allowed them to realize that some specifics of the credit portfolios and macroeconomic environments may reveal

quite stable along time. For those cases, a model built with a static learning setting may seem appropriate, if used during stable phases.

1.2. How industry currently handles credit scoring models rebuilds?

Developing and implementing a credit scoring model can be time and resource consuming, easily ranging from 9 to 18 months, from data extraction until deployment. Not infrequently banks use unchanged credit scoring models for several years. Since models are built using a sample file frequently comprising two or more years of historical data, in the best case scenario, data used in the models are shifted three years away from the point they will be used. If conditions remain unchanged, then this does not significantly affect the accuracy of the models. Otherwise, models' performance can greatly deteriorate over time. The recent financial crisis confirmed that financial environment greatly fluctuates, in an unexpected manner, posing renewed attention regarding models built upon timeframes that are by far outdated. By the crisis, many financial institutions were using stale credit scoring models built with historical data of 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 with a great deal of empirical evidence (Sousa et al., 2015, Rajan et al., 2015, Lucas, 2004, Avery et al., 2004).

Before the IRB approach has been introduced, in the Basel II Accord, financial industry was less motivated to rebuild credit scoring models. Then, financial institutions often outsourced the models development to analytics providers, while assigning some internal staff to these activities. Changes to the models of self-initiative were scarce, because it was expensive and time-consuming to build new models. Nowadays, part of the banks using the IRB approach internalized this activity, because they are required to closely monitor the performance of the models and suitably respond to changes. Not infrequently, this entails multiple local adjustments in the models to improve their accuracy, which may be as costly and time-consuming as developing a new model from scratch. The process of developing a new model or locally adjusting models factors can largely depend on judgmental reasoning or over-layered decision frameworks. Given that the time to decide or adapt may take too long to occur, the aim of opportunely adjusting rating systems, or credit policies, may become disappointing. Models' adjustments, or calibration, commonly consider selected macroeconomic public indicators, and should be opportunely revised. In so doing, resulting adjusted scores translate a combination of the customers' specific risk with systemic risk. 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 regulators define some rules, they are rarely convergent; and different countries favour different calibration choices (EBA, 2013).

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. The behaviour of the individuals and their ability of repaying their debts change when the economic conditions evolve in the economic cycle. In addition, 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., 2015). In order to adapt 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 learn adaptively (Gama et al., 2014). Advances on the concepts for knowledge discovery from data streams suggest novel perspectives to identify, understand and efficiently manage dynamics of behaviour in consumer credit in changing environments. In a world where the events are not preordained and little is certain, what we do in the present affects how events unfold in unexpected ways. Implementing a dynamic modelling framework for credit risk assessment enables monitoring the

existing credit risk models and timely anticipating potential models improvements in response to a change. Furthermore, it makes possible exploring different amounts of memory concurrently. The enormous advances in the processing power and in storage capacity, together with the advances in the streaming analytics suggest the practicality of implementing an adaptive modelling framework. Some experts believe that regulators are unlike to approve models whose coefficients change over time. Under these circumstances, a model can be used for years without further relearning throughout its utilization, possibly for several years, independently of the changes. This research provides new evidence on the significant degradation of credit scoring models based on static learning, broadly used among academics and practitioners.

This paper follows in section 2 with a brief description of the settings and concepts of the supervised learning problem and score formulation. It also presents the fundamental ideas of adaptive learning. In section 3 we present the conditions behind our case study, by providing an overview of the Freddie Mac's database and of the main dynamics over the period 1999-2013(Q1). Section 4 presents the adaptive modelling framework used in our experimental design. Section 5 presents the results. Firstly, we compare the performance of the models developed with the adaptive learning procedures versus a baseline static model. Secondly, we contrast the results of the STM with the LTM configuration. Conclusions and ideas for future work using the dynamic modelling framework are traced in section 6.

2. Settings and concepts

In this work we import some of the emerging ideas in concept drift adaptation into credit risk assessment models (Adams et al., 2010, Pavlidis et al., 2012). This is a field of research that has been receiving much attention in machine learning over the last decade, as an answer for suitably shaping the models and processes to a reality that is ever-changing over contexts and time. The settings and definitions adopted in this paper replicate the general nomenclature surveyed by Thomas (2010).

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 training examples – train 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), support vector machines (SVM), 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 – validation or out-of-sample data sets. The idea is to anticipate the accuracy of that function in future predictions of new examples where $\mathbf{x}$ 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(**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:

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

where B represents the bad class. Since p(G|x)+p(B|x)=1, the probability of the complementary class comes as

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

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

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

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

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

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 work. The basics of credit scoring and the most common approaches to build a scorecard 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).

2.2. Methods for adaptation

Traditional methods for building a credit scoring model consider a static learning setting. In so doing, this task is based in the learning in a predefined sample of past examples and then used to predict new examples, an actual or a potential borrower, in the future. This is an offline learning procedure, because the whole training data set must be available when building the model. The model can be used for prediction only after having completed the training, 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 for 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.

In Finance, it remains unresolved whether it is best having a long memory or forgetting old events. If on the one hand, a long-term memory (LTM) is desirable because it allows recalling a wider range of different occurrences, in the other, many of those occurrences may no longer adjust to the current 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 score modelling, this has been indirectly discussed by practitioners and researchers when trying to figure out 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 and reliable 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.

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 flow of data for dynamic modelling.

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 data set 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.

Pavidlis, Tasoulis, Adams and Hand (2012) were the first to propose an adaptive online algorithm for logistic regression in the classification of credit applications. According to the authors, this methodology has the potential to adapt to population drift. 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 data sets 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 data set consisting of 92,258 UPL applications accepted between 1 January 1993 and 30 November 1997 in the United Kingdom, revealed that the model was unable to outperform a static classifier built with the data of the beginning of the period, 1993. The authors provide insufficient thoughts regarding this finding, regardless of the existence of population drift in the data set, which had been documented in a previous study of Kelly, Hand and Adams (1999). Our paper is the first to document the dominance of the adaptive over static modeling frameworks in a real-world financial data set.

3. Case study

Our research was conducted in the Freddie Mac's single family mortgage loan-level database, first published in March 2013. We follow the performance of 16.7 million of fully amortized 30-year fixed-rate mortgages loans in the U.S., originated between January 1, 1999 and March 31, 2013. Disseminating these 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 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 data set is a living data set updated over time, typically at the end of each quarter, with the origination and performance data being summarized by month, from the origination point until the most recent reporting period.

3.1. Origination data

Our research considers a set of information that was available to the lenders at the time of the of the mortgage origination point, which is further detailed in Table 1. The release changes of the database are published online as well as a general user guide describing the full file layout and data dictionary (Freddie Mac, December 2013). 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 2014).

Table 1: Data available to the lenders at the time of the origination.

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 (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. 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). Ratios falling outside the range 6% and 105%, are disclosed as unknown.	Numeric
Original debt to income (DTI) ratio	Debt to income ratio is based on the following calculation: *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; *Income*: the total monthly income used to underwrite as of the date of the origination of the mortgage loan. Ratios greater than 65% or unknown are passed as null values. Note: The disclosure of the data set 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 data set.	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 data set.	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

3.2. Performance data

The loans performance[2] is outlined in a monthly basis. At the time of this research, data for performing loans and those that were up to 180 days delinquent were available through June 30, 2013. For each loan, there is a complete monthly historical report of the debt service, since the time of the origination until the most recent reporting period containing the:

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

Loan delinquency status - a value corresponding to the number of days the borrower is delinquent, based on the due date of last paid instalment reported by servicers to Freddie Mac,

[2] 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. 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.

calculated under the Mortgage Bankers Association (MBA) method. A code indicating the reason the loan's balance was reduced to zero if the loan has been:

- Prepaid or matured (voluntary payoff);
- Foreclosure (short sale, third party sale, charge off or note sale);
- Repurchase prior to property disposition, or;
- Real-estate Owned (REO) disposition.

For assigning the default event target we considered that a borrower entered in default if he was ever 90 or more days delinquent, the typical definition used under the Basel II.

3.3. Descriptive analysis

3.3.1. Lending before and after the crisis

The evolution of new loans over the analysed period illustrates the U.S. housing bubble between 2001 and 2005. Higher peaks occur between 2001 and 2003, where the numbers of new loans continuously rose from nearly 800 thousand new loans in 2000 to 1,930 thousands in 2003 (Table 2, 1st row). This massive increase was one of the sources of the raise in the real estate property values that reached a peak by 2005. Loans underwritten between 2001 and 2005 account for 42% of the amount originated in the analysed period).

Table 2: Main indicators for loans originated between 1st January 1999 and the 1st quarter of 2013. Source: Freddie Mac.

Indicator	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 U.S. $)	138	104	260	262	311	188	240	202	202	210	345	177	131	192	58
Avg original loan amount ('000 U.S. $)	126	132	148	156	161	167	181	187	189	213	228	224	236	244	237
Scores concentration index[3]	13%	13%	13%	13%	14%	13%	13%	12%	12%	14%	20%	19%	20%	21%	20%
Scores stability index[4]		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
Interest rate (%)[5]	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

Loans' average rate was maintained through 2001 and 2002, both in the aggregate level and in the score in each bucket (Table 3). 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 cut short-term interest rates from 6.5% to 1%. The mortgage interest rates continued to decline until 2005. As the mortgage rates are typically set in relation to 10-year Treasury bond yields, this was an outcome of very low Fed funds' rates in the period.

[3] We used Herfindahl-Hirschman Index (HHI) to evaluate the scores concentration, for which values below 20% are commonly considered acceptable.
[4] We used population stability index, for which values bellow 0,25 are commonly considered normal.
[5] Calculated as the weighted average of rates by score buckets.

Table 3: Average original interest rate. Loans originated in the period 1999-2013(Q1). Unit: %. Source: Freddie Mac.

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

3.3.2. Tracking the dynamics

3.3.2.1. Lending by score

In the U.S., borrowers' score is a key indicator in the mortgage lending. From 2009 onwards, the amount on loans in the scores below 620 is zero (Table 4), meaning that low scores' borrowers were firmly contained since then. By that year, lending moved markedly to the highest scores (see the shape of the histogram moving from year 2008 to 2009, in Tables 4 and 5). This effect is also captured in the score stability index that jumps suddenly from 0.10 to 0.28 in 2009 (Table 2, row 5). As a consequence, there is an increase in the concentration by scores from 14% in 2008 to more than 20% since 2009 (Table 2, row 4). The number and amount of loans diminished after 2009.

Table 4: Total number of loans by score in the year. Loans originated in the period 1999-2013(Q1). Source: Freddie Mac.

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	13	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	860
[650;675[		122	85	183	168	178	125	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	25	2.390
[725;750[		195	125	278	262	313	172	186	148	139	145	210	106	74	104	34	2.997
[750;775[		193	135	309	303	379	197	214	167	162	182	328	162	116	166	53	3.666
[775;800[		98	89	225	240	315	160	212	172	170	211	470	240	172	250	76	3.000
[800;850]	Lowest risk	11	13	32	37	52	30	77	71	74	91	235	130	91	142	41	1.125
Global		1.083	787	1.757	1.683	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 5: Amount granted in the year. Loans originated in the period 1999-2013(Q1). Unit: U.S. B$. Source: Freddie Mac.

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	530
[775;800[		32	11	32	37	49	26	39	33	34	47	132	56	43	64	19	825
[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%

3.3.2.2. Default by score

The vintage cumulative default rates 1 year after the loans have been originated are shown in Table 6. Real default rates are extremely irregular over time and real default concept drift is noticeable through 1999 to 2011. Although the mortgage bubble had expanded between 2001 and 2005, the higher loans default rates occurred for the loans originated in the period 2004-2008. Loans originated in the beginning of the boom do not have higher defaults than the loans in the precedent years; only the borrowers that have underwritten after 2003 reached higher default rates. Our analysis also show 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 (Sousa et al., 2015). 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.

Table 6: Cumulative default rates 1 year after the loan were originated. Unit: As % of the loans in the score bucket.

Score bucket	Credit risk assessment	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
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%
[300;550[	Highest risk	2,96%	4,62%	6,02%	6,83%	2,72%	3,80%	4,01%	3,92%	6,67%	12,87%	6,82%	7,14%	0,00%
[550;575[	↑	1,73%	3,23%	3,47%	4,06%	1,99%	2,68%	2,16%	3,31%	7,10%	10,32%	2,99%	3,13%	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%
[600;625[		1,05%	2,42%	2,52%	2,26%	1,12%	1,17%	1,35%	1,89%	3,37%	6,17%	1,93%	1,01%	0,71%
[625;650[		0,75%	1,42%	1,31%	1,29%	0,65%	0,77%	0,89%	1,16%	2,32%	5,22%	1,91%	0,78%	0,67%
[650;675[		0,48%	0,80%	0,70%	0,70%	0,39%	0,44%	0,60%	0,72%	1,63%	3,37%	0,91%	0,54%	0,39%
[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%
[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%
[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%
[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%
[775;800[		0,06%	0,08%	0,07%	0,07%	0,03%	0,03%	0,08%	0,06%	0,17%	0,25%	0,03%	0,02%	0,03%
[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%
Global		0,28%	0,56%	0,48%	0,48%	0,19%	0,27%	0,32%	0,41%	1,02%	1,35%	0,15%	0,08%	0,07%

4. Modelling framework

4.1. Adaptive modelling

The dynamic modelling framework implemented in this research considers that data is processed batch-by-batch, as illustrated in Fig.1. Sequentially, at each year, a new model is learned from a previous selected window, including the most recent year. To represent the time evolution, we assumed that the current year gradually shifts from 1999 until 2011.

Each learning unit of the model was grounded on a static setting. At each year, instances for modelling are selected from all previous available batches, according to a selection mechanism. We use instance selection methods to test the hypothesis under investigation. Two methods for tackling default concept drift were implemented – a long-term memory and a short-term memory (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.1(a)). The process is incremental, so every time a new instance arises, it is added to the training set, and a new model is build. This schema should be appropriate to detect mild concept drifts, but it is unable to rapidly adapt to major changes. Models of this schema should perform suitably in stable environments. A shortcoming of this incremental schema is that the training data set quickly expands which may require a huge storage capacity.

In the STM windowing configuration, the model development uses the most recent window. With this schema, Fig.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 concept. Under this setting, the dynamic modelling should quickly adapt to changes. The most extreme case of short memory time window is when only the current example is considered to train the new model, which represents to the online learning without any memory of the past. A deficiency of this method is that it often lacks of generalization ability in stable conditions, which is amplified with extremely short windows.

(a) LTM. (b) STM.

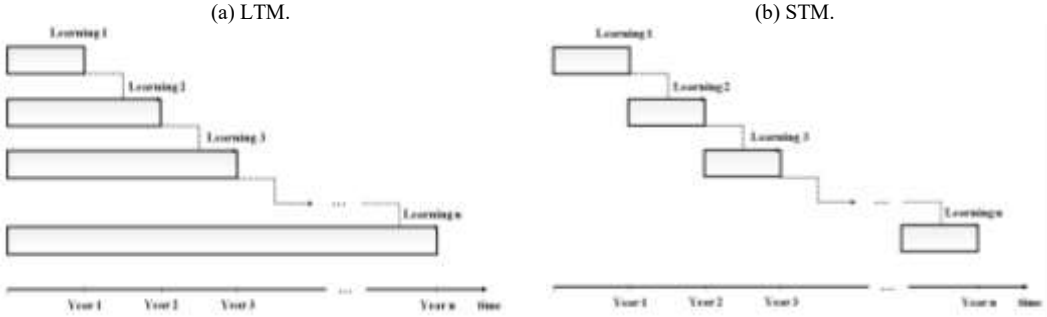


Fig. 1. Adaptive learning windowing configurations.

The research questions of this study should be answered following the reasoning:

- If the LTM outperforms the STM, then more recent data are not fundamental for the prediction; the environment of the decision-making should be in a stable phase. Otherwise, the default concept is drifting, and so the most recent data are more relevant for the prediction.
- If a model built with static learning in the first window of the period has the best performance, then older data can improve the prediction. This may happen, for example, when a new credit product is launched, and the credit decision-making criteria are adjusted afterwards. In such case, the oldest data are more representative, as they can illustrate a more diverse range of risk behaviours.

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, on their 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. First, 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 using the weight of evidence (WoE). Variables were transformed using the WoE in the complete training data set

$$\mathrm{WoE} = \ln\left(\frac{\mathrm{g/G}}{\mathrm{b/B}}\right), \tag{5}$$

where g and b are respectively the number of good and the number of bad in the bin, and G and B are respectively the total number of good and bad in the population sample. The larger the WoE the higher is the proportion of good customers in the bin. Numerical variables were firstly binned. Cases where the calculation of the WoE rendered impossible, i.e. one of the classes without

examples, were given an average value. The same principle was applied to values out of the expected ranges. The strength of each potential characteristic was measured using the information value (IV) in the training data set

$$\mathrm{IV} = \sum_{i=1}^{n}\left({}^{g}/_{G} - {}^{b}/_{B}\right)\mathrm{WoE}, \tag{6}$$

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 process (Silva and Cardoso, 2015).

The scorecard design was wrapped in a forward feature selection process to find the optimal subset of characteristics. The selection process stops when no other characteristic added significant contribution to the information value (IV) of the model. In this application the threshold was set for a minimum increment of 0.03. Third, the performance of the model is measured with the Gini coefficient, equivalent to consider the area under the ROC curve (AUC). This is a typical evaluation criterion among researchers and in the industry. 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.

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 originated in the development year, and the set of loans originated in the following year, an out-of-sample performance.

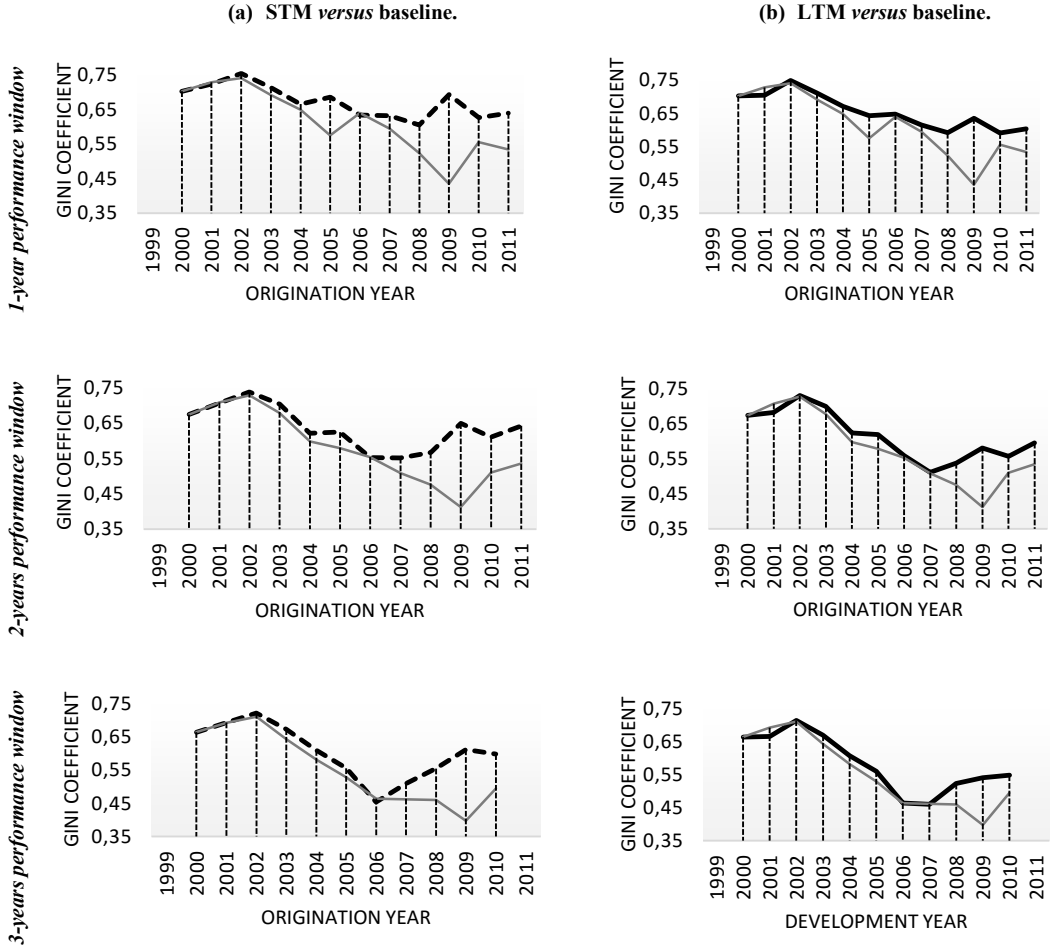
The vintage curves presented in a previous study of Landy, Ashworth and Yang (2014) suggest that the cumulative default rates of this portfolio reach a plateau by the fifth year. Since most of the default events occur between the first and the fifth year after the loan has been originated, we assumed that the performance measures of the models should be calculated within this timeframe. So, despite that the models' learning has considered a fixed target concept, a borrower being ever 90+ days delinquent in the year after the loan has been originated, the performance was measured in five annually-incremental performance windows, from a 1-year performance to a 5-years performance window, where a borrower is assigned to the default class if he was ever 90+ days delinquent within that timeframe. In so doing, our aim is to realize the true performance of the models over the most relevant part of the life of the asset, rather than just illustrating the 1-year performance as conventional approaches usually do. The last origination year for the performance measurements vary according to the length of the performance window (e.g. for the loans underwritten in 2009, only a 4-years performance can be measured until 2013, and for the loans underwritten in 2012, only a 1-year performance can be measured until 2013). Hence, the 5-years performance is measured until the origination year 2008, the 4-years performance is measured until 2009, the 3-years performance until 2010 and the 1 and 2-years performance until 2011. It is worth noticing that in the last origination year, the performance window length is rounded up in a half-year, because at the time of this study, loans' performance was available only until June 2013. For a similar reason, the 1-year performance is not presented for the loans originated in 2012, since the performance of the loans originated in December could only be measured through half-year performance window, deemed insufficient.

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

5.1. Adaptive learning versus baseline static learning model

A baseline static model was developed using the loans originated in the first year of the period – 1999. This model was applied over the entire period, i.e. to each loan originated between 2000 and 2011, and the performance assessed in each year, for the five performance windows lengths. Results are presented in Fig. 2, where the performance of the adaptive learning, in the STM and LTM configurations, is compared with the performance of the baseline static model. For more realism, the results of the adaptive learning procedure consider that a model is applied to the loans originated in the year following the year used to learn the model. In fact, for a complete realism, a minimum 2-years window should be used to have a 1-year performance window for all the observations. We have chosen not to apply this principle because we would have to discard the performance for 2000, at the beginning of the housing bubble, which performance we are interested to capture. In fact, considering the huge volume of available data, the learning could be based in 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 degrades over time, which intuition also suggests. When compared with the adaptive learning procedure, the extent of degradation decays more significantly from 2007 onwards and most noticeably in the aftermath of the Crisis, in 2009. This finding is consistent for every performance window length.



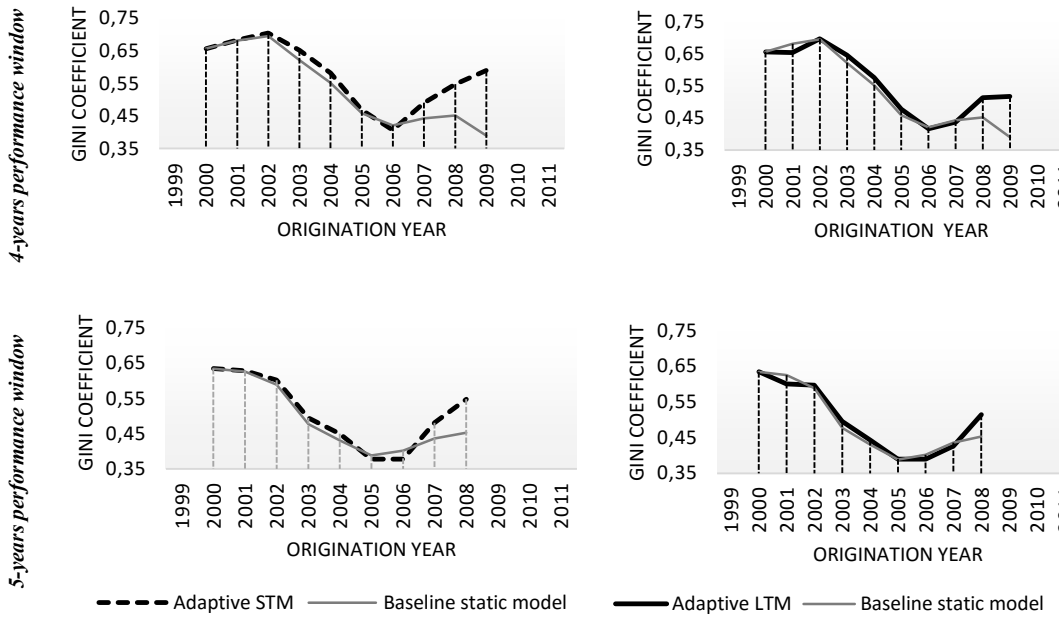
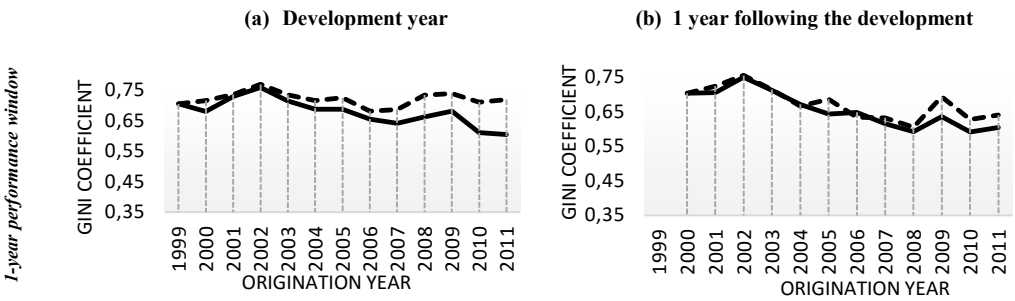


Fig. 2. Adaptive learning model *vs.* baseline static model; model applied to the loans originated 1 year after the development.

5.2. Adaptive short-term memory (STM) versus adaptive long-term memory (LTM)

By comparing the performance in the short-term memory (STM) with the long-term memory (LTM) configuration, in Fig. 3, we find that the STM configuration consistently outperforms the LTM. This finding is consistent both in the development test sample, here referred as the development year, 1 year following the development. As it has been anticipated, the STM configuration consistently produced the highest performance during periods of exacerbated financial distress, from 2007 onwards. Although we had conjectured otherwise, the results of our analysis did not provide evidence on that the LTM outperforms the STM in stable conditions. We speculate that the length of the memory used in the STM configuration was not sufficiently short, carrying a quite suitably performance along the analyzed period, not only during unstable phases but also in stable conditions. This hypothesis also suggests that the STM performance could have been further improved if we have tried shorter-term configurations. However, it is worth noticing that smaller windows may be harder to have acceptance by the industry, considering cost, business and regulatory constraints.



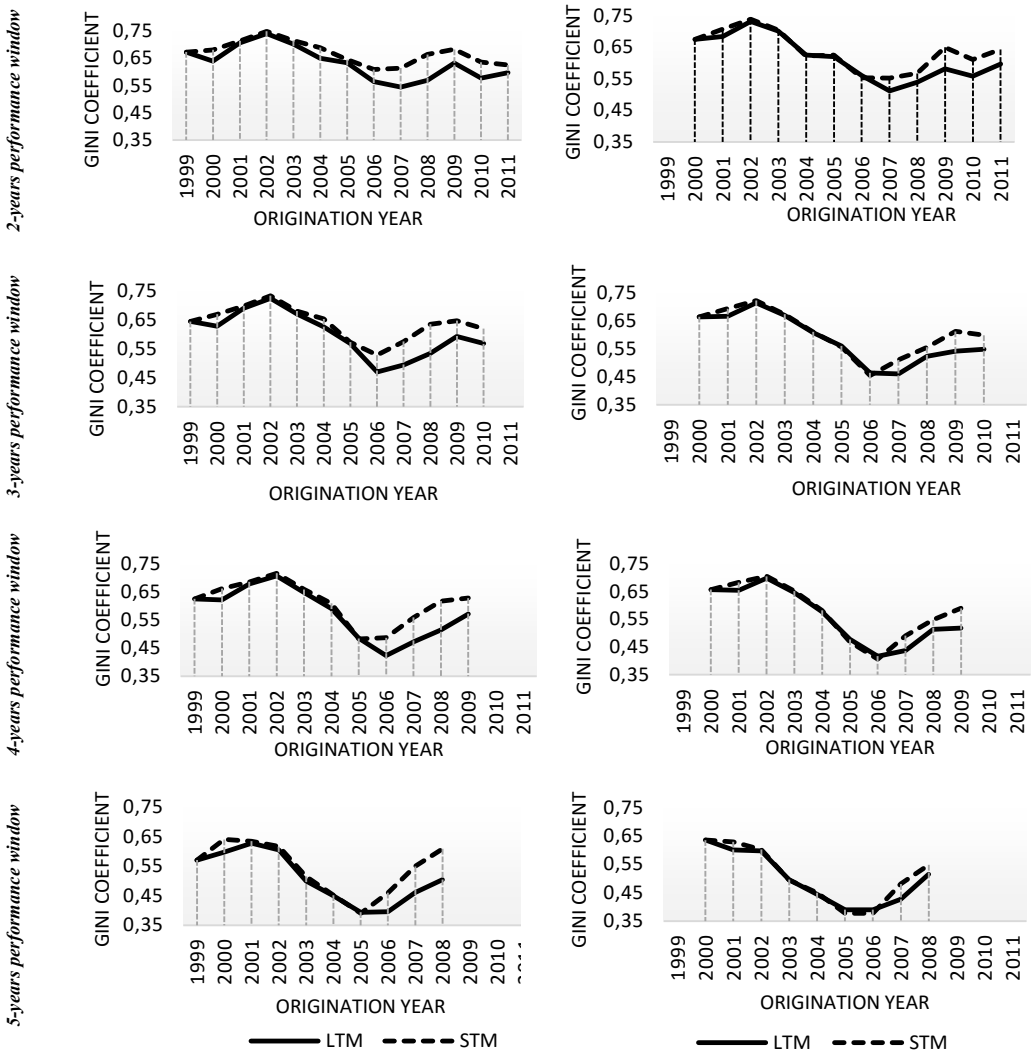


Fig. 3. Performance of the models built with the adaptive learning framework in the two memory configurations.

6. Conclusion

Credit risk assessment is one area where data mining and forecasting tools have largely expanded over the last 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 so, they are quite insensitive to changes, like population drifts 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 the drift detection, and the occurrence of a drift may suggest eventual corrective actions 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 model learning ability and models' predictive power through time in real evolving landscapes. Central to our study is the idea that model learning is improved when acting similarly to human learning based on the 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 learnt from a real-world data set of 16.7 million loans that were at the epicentre of the global crisis, the Freddie Mac's single family mortgage loan-level data set, 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 improve forecasting accuracy, and STM allows a quick adaptation to changes. Older information did not improve forecasting accuracy, but no general rule can be extracted, since this 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 conjecture that this may have been a consequence of having used an insufficiently short window length in the STM configuration. Complementing with the previous studies, our paper presents renewed relevant empirical evidence on that traditional modelling frameworks significantly degrade over time and that models predictive effectiveness is largely improved when adaptive learning frameworks are applied.

There are some real business problems with rebuilding models over time. First, lenders have little incentive to enhance the existing rating systems frameworks because it is expensive and time-consuming to build new scorecards. They need to be internally tested and validated, and then regulators need to approve them. Second, regulators still promote models whose coefficients do not change over time. This is one area where this new thoughts could be encouraged. Furthermore, more sophistication is needed and more effort should be put to promote renewed principles in the risk assessment frameworks. 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. First, a smaller sample is required for relearning the model, while keeping memory of the past. Second, a model which depends on the previous assessments is more palatable, and hence, it is more likely to be accepted. Another working track is developing 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 the adaptation ability.

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*, OUP Oxford.

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.

BCBS 2006. International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version. *Bank for International Settlements*.

BIS 2004. Implementation of Basel II: Practical Considerations. *Bank for International Settlements*.

EBA 2013. 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.

FREDDIE MAC December 2013. Single Family Loan-Level Data set General User Guide *In:* MAC, F. (ed.). Freddie Mac.

FREDDIE MAC June 2014. Single Family Loan-Level Data set - Summary Statistics. Freddie Mac.

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.

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

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*.

KELLY, M. G., HAND, D. J. & ADAMS, N. M. Year. The impact of changing populations on classifier performance. *In:* Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999. ACM, 367-371.

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

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

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

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

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

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.

SILVA, F. B. S. & CARDOSO, J. S. 2015. Differential Scorecards for Binary and Ordinal data. *Intelligent data analysis, forthcoming*.

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. University of Porto, Portugal, School of Economics and Management.

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

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.

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