Dynamic credit score modeling with short-term and long-term memories: The case of Freddie Mac's database

Article *in* Journal of Risk Model Validation · March 2016

DOI: 10.21314/JRMV.2016.147

Maria Sousa
University of Porto
6 PUBLICATIONS **10** CITATIONS
SEE PROFILE

João Gama
University of Porto
361 PUBLICATIONS **6,096** CITATIONS
SEE PROFILE

Elísio Brandão
University of Porto
40 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Dynamic credit score modelling with short-term and long-term memories: the case of 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 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.

Keywords: Freddie Mac, score, PD, memory, credit risk, scoring, adaptive learning

JEL: C4, C6, C8, C9

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

[1] Corresponding author

Email addresses: 100427011@fep.up.pt (Maria Rocha Sousa), jgama@fep.up.pt (João Gama), ebrandao@fep.up.pt (Elísio Brandão) 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 data sets. 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 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 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 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 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 data set, which had been documented in a previous study of Kelly, Hand and Adams (1999). In a previous paper (Sousa et al., 2016), we have already put forward an adaptive model for credit scoring. However, we also lacked a proper experimental validation.

The present paper is the first to document the dominance of the adaptive over static modelling frameworks in a real-world relevant financial data set, the Freddie Mac's database.

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, 2013).

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.

1.2. Structure of the paper

Section 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 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 4 will present the adaptive modelling framework used in our experimental design. In section 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 6.

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 data set 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 data set 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.

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 data set 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.

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 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 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

Name	Short description	Type
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

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.

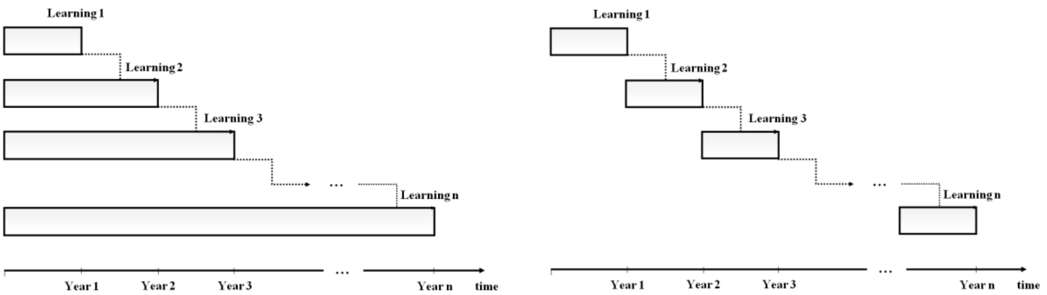
We consider that a borrower defaulted if he was, at any point, 90 or more (90+) days delinquent, the typical definition used under Basel II. Later, in section 5, we will describe the construction of scorecards based on a supervised learning procedure and a binary target, where a borrower is assigned to the “bad” class, if he defaulted, and is assigned to the “good” class otherwise.

4. Adaptive modelling framework

The dynamic modelling framework implemented in this research considers that data is processed batch-by-batch, as illustrated in Fig.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.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 data set 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.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. 1. Adaptive learning windowing configurations.

5. 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

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

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

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

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 2 and 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 2: Marginal contribution of the variables in each learning unit of the LTM configuration.

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 3: Marginal contribution of the variables in each learning unit of the STM configuration.

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	2,244	2,548	2,957	3,119	2,359	2,563	2,922	2,922	2,2312	2,803	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.

6. 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.

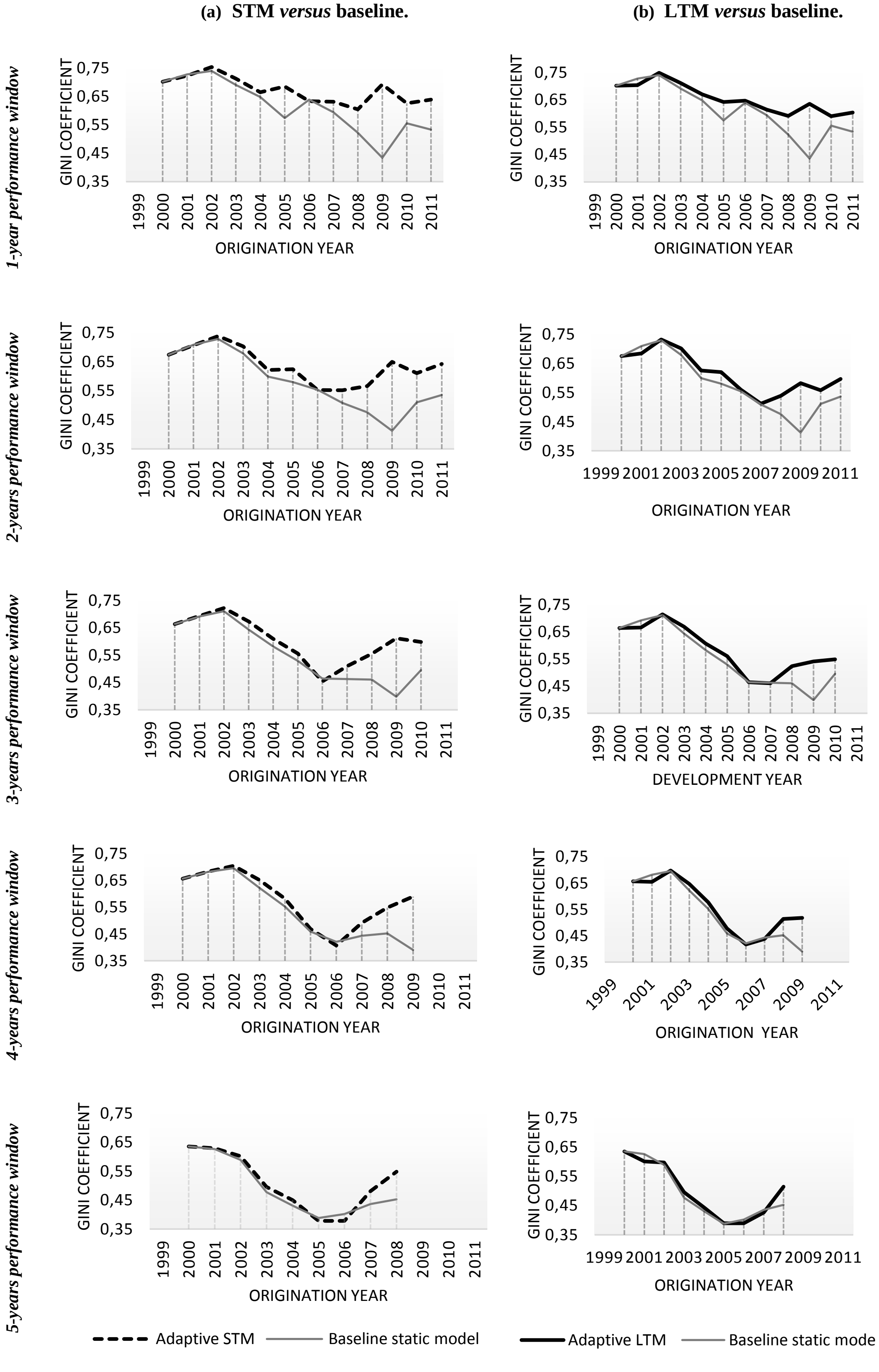
The vintage curves presented in a previous study of Landy, Ashworth and Yang (2014) suggest that the cumulative default rates of this portfolio reached a plateau by the fifth year. Since most of the default events occurred between the first and the fifth year after the loan had been granted, we assumed 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.

6.1. 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. 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.

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

6.2. 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. 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.

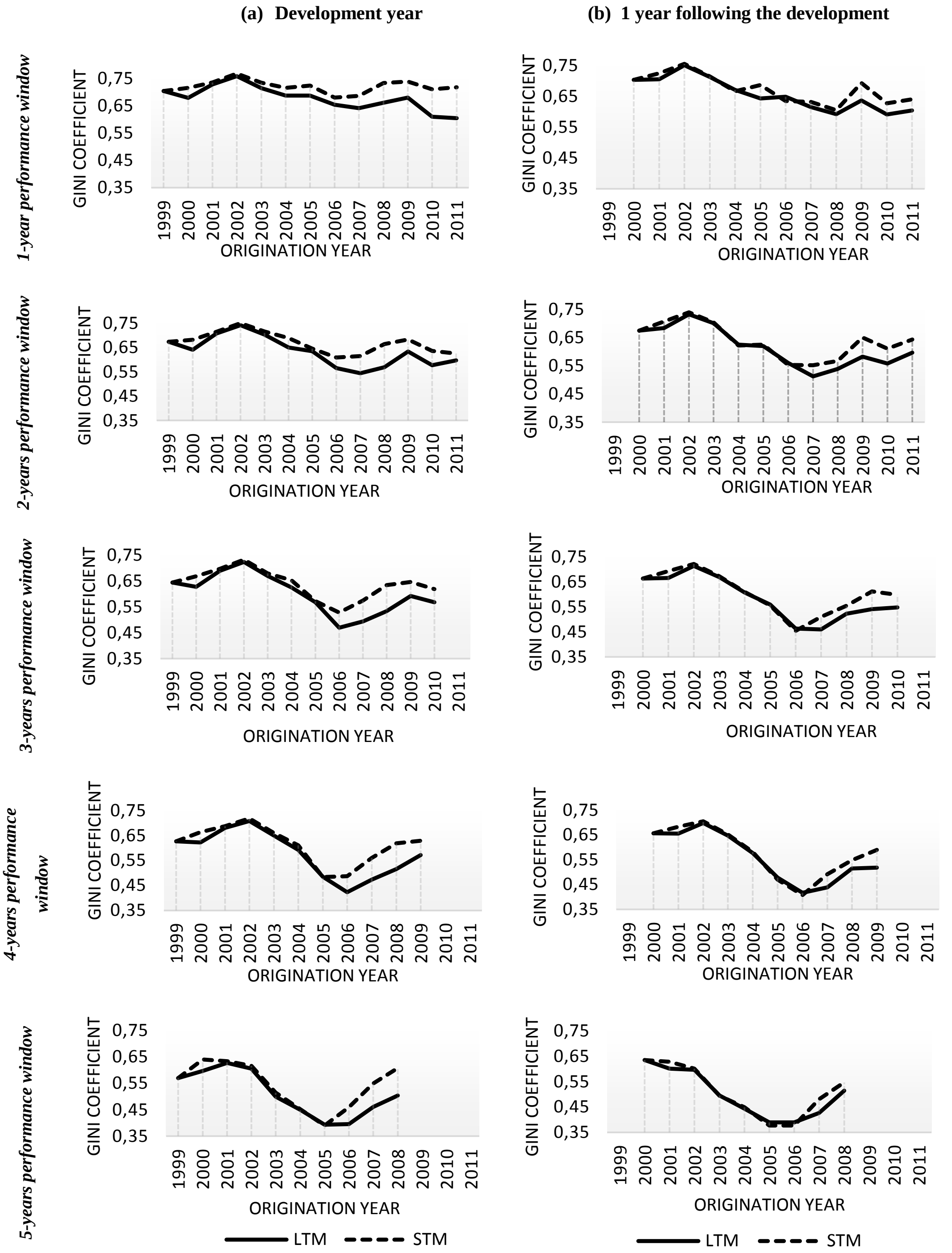


Fig. 3. Performance of the models built with the adaptive learning framework in the two memory configurations.

7. Conclusion

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 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 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 paper 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.

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.

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.

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.

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

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

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

FREDDIE MAC June 2014. Single Family Loan-Level Dataset - 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. 1999. The impact of changing populations on classifier performance. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.

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.

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.

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.

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.

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