Estimation techniques for deriving the Basel and IFRS 9 LGD estimates on retail bank portfolios

M Joubert

orcid.org 0000-0002-4711-6145

Thesis submitted in fulfilment of the requirements for the degree *Doctor of Philosophy in Risk Analysis* at the North-West University

Promoter: Prof H Raubenheimer

Co-promoter: Prof T Verster

Graduation May 2019

26801213

Contents

Abstract

A stable financial system is essential for growth in banks. A financial crisis can damage banks, as was seen in the financial crisis of 2008. Banks are subject to government regulation to reduce the risk of future financial crises. Amongst several requirements, capital requirements, as set out by local government, are influenced by the Bank for International Settlements' Basel Committee on Banking Supervision. The requirements, as set out in the Basel Accord, allow banks to build risk models for three risk drivers, namely the probability of default (PD), loss given default (LGD) and exposure at default (EAD). The risk drivers are combined to predict the unexpected credit loss that is used as a safety cushion against unexpected credit losses. Banks are also subject to financial reporting and disclosure requirements. International Financial Reporting Standards (IFRS) are standards issued by the IFRS Foundation and the International Accounting Standards Board (IASB). The IFRS 9 standard gives guidance with regard to the estimation of impairments and typically the same three risk drivers are used. Impairments models are used to estimate provisions that banks need to hold against expected credit losses. The accuracy of these risk drivers are also key to the stability of banks. The objective of this thesis is to develop LGD models for Basel and IFRS 9 that adhere to the required regulations. LGD methodologies can be classified into direct and indirect methodologies. Under Basel, a direct and indirect LGD model was developed. The direct LGD model was adapted for IFRS 9 requirements.

Survival analysis is one of the approaches used in direct LGD modelling. A standard method in this approach is the EAD weighted survival analysis (denoted by EWSA). The first article will aim to enhance the survival analysis estimation of LGD. Firstly by using default weighted LGD estimates and incorporating negative cashflows and secondly by catering for over recoveries. We will denote this new method to predict LGD as the default weighted survival analysis (DWSA). These enhancements were motivated by the fact that the South African Reserve Bank requires banks to use default weighted LGD estimates in regulatory capital calculations. Therefore, by including this into the survival analysis approach, the model is aligned more closely to regulations. Recovery datasets used by banks include both negative and over recoveries. By including these into the LGD estimation, the models are more closely aligned to the actual data. The assumption is that the predictive power of the model should therefore be improved by adding these changes. The proposed model is tested on eight datasets. Three of these are actual retail bank datasets and five are simulated. The datasets used are representative of the data typically used in LGD estimations in the South African retail environment.

When the indirect LGD methodology is used, two components exist, namely the loss severity component and the probability component. Commonly used models to respectively predict the loss severity and the probability component are the haircut- and the logistic regression models. In the second article, survival analysis is proposed as an improvement to the more traditional logistic regression method. By testing the MSE (mean squared error), bias and variance of the two methodologies, it was shown that the improvement enhanced the model's predictive power. The proposed LGD methodology (using survival analysis) was applied on two simulated datasets and two retail bank datasets, and outperformed the logistic regression LGD methodology. Additional benefits included that the new methodology could allow for censoring as well as predicting probabilities over varying outcome periods.

The third article is aimed at adapting the DWSA method, used in the first article to model the Basel LGD to estimate the LGD for IFRS 9 impairment requirements. The DWSA methodology allows for over recoveries, default weighting and negative cashflows. This IFRS 9 LGD is used in the calculation of the expected credit losses (ECL) as per the IFRS 9 standard. The IFRS 9 LGD methodology that is described in this paper makes use of survival analysis to estimate the LGD. The Cox proportional hazards model allows that a baseline survival curve can be adjusted to produce survival curves for different segments of the portfolio. The forward-looking LGD values are adjusted for different macro-economic scenarios and an ECL is calculated for each scenario. These ECL values are probability-weighted to produce a single ECL number. This paper illustrates the IFRS 9 LGD as well as the ECL on a real dataset from a retail portfolio of a South African bank.

Key words: Loss Given Default, Survival Analysis, Basel, IFRS 9, Retail Credit

Acknowledgements

Embarking on further studies is never a decision to be taken lightly. It requires significant effort, hours and dedication. Standing at the end of the journey, it is easy to see that it was indeed worth the effort. For me, the difference between quitting and continuing with a project lies in your motivation. Key to my motivation were people from different walks of life, each playing an important role in keeping me on track. Some assisted in creating the right environment to further my studies, whilst others assisted on a more technical basis. As much as this document constitutes my own work, there have been many influencing individuals that impacted the direction as well as the quality of the work. It is therefore these individuals that I would like to acknowledge and thank in this section:

- I would like to express my sincere gratitude to my supervisor, Helgard Raubenheimer, for the continued support of my PhD study, for his patience, motivation and immense knowledge. As one of the co-authors of three articles in this document, I would like to thank you for your insightful comments and suggestions. Thank you for mentoring me, for encouraging my research and for allowing me to grow as a research scientist.

- To my co-supervisor, Tanja Verster - thank you for the time spent reviewing and providing valuable insights. Your suggestions regarding the topic as well as advice on structuring the document did not go unnoticed. You provided the much needed "golden thread" that connected the various elements. As a co-author to the three articles that constitute the thesis, thank you for the time spent perfecting our research and invaluable editing skills.

- A special thanks to my family. Words cannot express how grateful I am to my wife, Esther, my father, Fred, my mother, Elmien, and my daughter, Mia, for all the sacrifices that you've made on my behalf. Your prayers sustained me thus far.

- Above all, I owe it all to Almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to its completion.

Preface

The article format was chosen for this thesis. The research reported in this thesis was done in conjunction with my supervisor, Professor Helgard Raubenheimer, and co-supervisor, Professor Tanja Verster. The articles were written for the purpose of this thesis and were submitted to the indicated journals for publication. The co-authors provided their permission that these articles can be submitted for degree purposes. I was the main author for these articles, my promoter and co-promoter reviewed the articles on a regular basis and made suggestions for changes. These articles are:

- Title: Default Weighted Survival Analysis to directly model Loss Given Default. Authors: M. Joubert, T. Verster and H. Raubenheimer. The article was published in the South African Statistical Journal 2018, Vol. 52, No. 2, 173-202. Chapter 2 of this thesis consists of this article.

- Title: Making use of Survival Analysis to indirectly model Loss Given Default. Authors: M. Joubert, T. Verster and H. Raubenheimer. This article was accepted for publication in the Operations Research Society of South Africa (ORSSA) (2018). Chapter 3 of this thesis consists of this article.

- Title: Adapting the Default Weighted Survival Analysis Modelling approach to model the IFRS 9 LGD. Authors: M. Joubert, T. Verster and H. Raubenheimer. This article was submitted to the Journal of Empirical Finance (2018). Chapter 4 of this thesis consists of this article.

The literature study and motivation in Chapter 1 is followed by Chapter 2, Chapter 3 and Chapter 4, which contain the above-mentioned articles. The downturn Basel Loss Given Default is modelled for secured and unsecured retail portfolios in Chapter 3 and Chapter 4, respectively. The IFRS 9 LGD is the focus of Chapter 4, and Chapter 5 concludes.

Chapter 1

Introduction

Contents of Chapter 1

List of Figures

List of Tables

INTRODUCTION

Key words: Loss Given Default, Survival Analysis, Basel, IFRS 9, Retail Credit.

Credit risk is defined as the risk or probability that a counterparty will default, due to failure to pay its credit obligations, in accordance with agreed terms. If this credit risk realises, an economic loss (shortfall) may be incurred should the bank not recover all monies due. Since the financial crisis of 2008, credit risk modelling has attracted a lot of attention. There is now a greater awareness of how the quality of credit risk models affects the amount of capital and impairments that banks are to keep. The development of robust and accurate credit risk models has become vital. The accurate estimation of credit risk will result in a competitive advantage for banks. The Basel Accord (BCBS, 2006) allows for banks to derive their own internal credit risk models under the advanced internal rating based (AIRB) approach. The Accord further allows banks to build risk models for three risk parameters, namely the probability of default (PD), loss given default (LGD) and exposure at default (EAD). The risk drivers are combined to estimate the capital that is used as a safety buffer against unexpected credit losses. Apart from keeping capital for unexpected loss, banks also need to hold provision for expected loss. Impairment models are used to estimate these provisions. The IFRS 9 standard (IFRS, 2014) gives guidance with regard to the estimation of impairments.

This chapter starts out by giving background and motivation (Section 1) pertaining to Basel and IFRS 9. This is followed by the thesis objectives in Section 2 and Section 3 contains a literature summary focusing on LGD modelling methodologies. Section 4 summarizes the chapter.

1. Background and motivation

The motivation of this thesis will start with an overview of Basel regulations and a synopsis of IFRS 9. The shortcomings identified will then form the rest of the motivation.

1.1. Basel

Prior to 1974, the non-existence of regulatory systems allowed for major disruptions of international financial markets. This led central bank governors of the G10 countries to establish a committee on banking supervision. The main aim of the Basel Committee on Banking Supervision (BCBS) was to first put measures in place to create financial stability through co-operation between its members. This was achieved by the improvement of quality of banking supervision worldwide (BCBS, 2015b, pp. 1).

The aim of the BCBS is to identify current or developing risks to the global financial system. This objective is achieved by the development of minimum standards as well as knowledge sharing of best practices for the regulation and supervision of banks. BCBS promote common understanding and sharing of information across country borders. The committee has no legal standing in any country. BCBS depend on national authorities to implement these standards and policies developed by them (BCBS, 2015b, pp. 1). The South African Reserve Bank (SARB) is the governing body that is responsible for regulating the South African banks which implement the Basel Accord.

The need for fundamental strengthening of the Basel II Accord was identified before the financial collapse of Lehman Brothers in September 2008. The financial crisis of 2008 was an eye opener for the banking sector as they were exposed to high leverage and low liquidity buffers. The banking sector was further exposed to inappropriate risk management and weak incentive structures. The combination of above-mentioned factors led to the mispricing of credit and liquidity risk and excess credit growth. The Basel committee issued "Principles of sound liquidity risk management and supervision" in September 2008 and followed in July 2009 with a further revised issue to strengthen the Basel II capital framework (BCBS, 2015b, pp. 4).

The Basel II Accord is structured into three pillars. The first pillar deals with minimal capital required, the second with the supervisory review process, and the third with market discipline. Credit risk, market risk and operational risk is covered in the Accord (BCBS, 2006, pp. 6).

Under pillar one, the Accord allows banks to calculate their capital for credit risk by following one of two approaches (BCBS, 2006): the standardized approach and the internal ratings-based approach. The standardized approach measures credit risk in a standardized manner. External credit ratings are used to determine the risk weight for certain exposures. The committee has also taken under consideration the possibility of introducing a standardized set of risk drivers and recognizes the challenges that exist with such an approach (BCBS, 2014, pp. 1). The internal ratings-based approach (IRB) is a more advanced approach where banks can develop and use internal models that require approval from the banks regulatory authority (BCBS, 2006, pp. 52). These models include the PD, LGD and EAD.

Under pillar two, a bank's management needs to have a process in place to assess if the capital held by the bank is consistent with their risk profile, and a strategy to maintain their capital levels (BCBS, 2006, pp. 205). Regulators need to review and evaluate this process and strategy and take appropriate action if banks are not compliant (BCBS, 2006, pp. 209). Regulators should expect banks to operate above the minimum capital levels and intervene at an early stage when required (BCBS, 2006, pp. 211–212).

Pillar three introduces a set of disclosure requirements that banks needs to adhere to. Regulators have many measures that they can require from banks. Certain of these measures will become compulsory (BCBS, 2006, pp. 226).

The IRB approach utilises the Asymptotic Single Risk Factor (ASRF) model. The ASRF model assumes that a borrower will default if the value of the borrower's assets falls below the value of its debts. Within the ASRF model, the distinction is drawn between idiosyncratic and systematic risk factors. The law of large numbers shows that idiosyncratic risk factors cancel each other out within a large portfolio consisting of small accounts. Systematic risk factors therefore are the remaining risk factor that needs to be considered. In this framework all systematic risk factors that affect borrowers similarly are assigned only one single risk factor. This risk factor represents the changing economic conditions, which has the same effect on all portfolios. Under the ASRF model, risk-weighted assets (RWA) are calculated using PD, LGD and EAD estimates. It is worth noting that despite guidance towards the ASRF model by the Basel committee, banks are given discretion to use the model that best suits their requirements when it comes to estimating and mitigating risk. The internal ratings-based approach (IRB) is a more advanced approach where banks can develop and use internal models (PD, LGD and EAD models) that require approval from the bank's regulatory authority (BCBS, 2006, pp. 52). The models must be accurate and predictive across the range of borrowers. Banks need to validate their models on a regular basis to ensure monitoring of performance and stability (BCBS, 2006, par. 417).

The PD comprises of a point in time (PIT) PD and a through the cycle (TTC) PD. A PIT PD is the probability of an account defaulting in the following year, as estimated at a particular point in time. The TTC PD is the probability that an account defaults over the economic cycle (a long-run average). When the LGD is developed over a downturn period (a period where a 'downturn' is observed in an economic cycle) it is referred to as the downturn LGD. The TTC PD and the downturn LGD is used as an input in the Basel RWA calculation.

The Basel II revision was issued during December 2010. It was named "Basel III: A global regulatory framework for more resilient banks and banking systems" (BCBS, July 2010). The three focus areas of Basel II were greatly improved (BCBS, 2015b, pp. 4): Basel III introduced stricter definitions of capital, higher minimum ratios and the introduction of a macroprudential. Basel II was a fundamental enhancement of the guidelines to the banking regulations worldwide. The Basel committee together with the group of twenty (G20) leaders emphasized the introduction of the reformed banking framework as defined by Basel III in such a way that it would not impede and disrupt the recovery of the real economy (BCBS, 2015b, pp. 5).

This thesis will focus on developing retail bank LGD models for credit risk by making use of the internal ratings-based approach that is allowed under pillar one of the Basel II Accord when modelling the unexpected loss for capital requirements. The expected losses are treated separately under IFRS 9 and is described in the following section.

1.2. IFRS 9

During 2005, the Financial Accounting Standard Board (FASB) and International Accounting Standard Board (IASB) began working on simplifying the reporting for financial instruments. The discussion paper, "Reducing complexity in Reporting Financial Instruments", was published during March 2008. The discussion paper identified several possibilities for improvement. These were supported by financial institutions. This resulted in the IASB adopting the project to its agenda during November 2008. In April 2009 IASB announced an accelerated timetable for replacing IAS 39. This was due to the financial crisis and the conclusion of the G20 leaders and the International Stability Board (IFRS, 2014, pp. 4).

IASB added to International Financial Reporting Standard 9 (IFRS 9) the requirement to account for expected credit loss on its financial assets. This requirement eliminates the threshold that was in IAS 39 for the recognition of credit losses. Under the impairment approach in IFRS 9, it is not necessary for a credit event to occur before credit losses are recognized. Expected credit losses and changes to such are reported. The amount in credit losses is updated at each reporting period, reflecting the change in credit risk. The result is more timely information in respect of expected credit losses (IFRS, 2014, pp. 6).

The International Accounting Standard Board published the new and complete IFRS 9 standard in the form of the document "IFRS 9 Financial Instruments" (IFRS, 2014). This document replaces most of the IAS 39 standard. Amendments were made to the classification and measurements of financial assets standards. It also includes new hedge accounting guidance. It contains new impairment requirements that will allow for earlier recognition of credit losses. According to this guideline, the financial statements of banks must reflect the IFRS 9 accounting standards for the period starting on 1 January 2018 (EBA, 2016, pp. 4).

The IAS 39 accounting standard makes use of provisions on incurred losses. Learning from the financial crisis is that expected losses, instead of incurred losses, should be used to calculate the provisioning for banks (GPPC, 2016, pp. 21). Under IFRS 9, a financial entity allows for expected credit losses. The expected credit losses should be equal to the lifetime expected credit losses, if the credit risk has risen significantly. When the converse is true, a financial entity may allow for credit losses equal to 12 month expected losses (IFRS, 2014, pp. 26).

An entity must assess the significance in change of credit risk (SICR) at each reporting date. When conducting this assessment, each entity should use the change in risk of a default occurring over the expected life of the instrument compared to the change in the amount of credit loss. For this assessment to be made, the risk needs to be determined at the reporting date and compared to the risk at the initiation date. This will be an indication of SICR since initiation. The initial presumption that a significant credit risk develops when a payment is 30 days overdue is not valid anymore as the credit risk will increase if an entity has enough substantial information to determine that there has been a significant increase in credit risk (IFRS, 2014, pp. 27).

The benefits of IFRS 9 compared to IAS 39 will be seen in the accounting model for financial instruments and on credit loss provisions. It will impact recognition of credit losses regarding the issue of raising insufficient provisions at too late a stage. It will improve the accounting recognition of loan loss provision due to a wider range of credit information that should be collected (EBA, 2016, pp. 4).

The expected credit loss model is a forward-looking model and should result in the early detection of credit losses. This will contribute to financial stability. IFRS 9 is expected to address regulating concerns. The expected credit loss module is aligned with existing regulating practices where credit institutions use an internal ratings-based (IRB) model which, requires calculation of expected credit losses rather than incurred credit losses when determining regulatory capital requirements (EBA, 2016, pp. 7).

The complexity of judgement that is required in the expected credit loss assessment could affect the consistent application of IFRS 9 across credit institutions. The comparability of financial institutions' financial statements will be impacted. The volatile nature of expected credit losses compared to incurred losses means that there will be more intensive oversight following implementation (EBA, 2016, pp. 7).

Most credit institutions have well established capital models for the measurement of unexpected losses; these models may be used as a starting point for estimating expected credit loss. Regulatory capital models are not suitable for the use of expected credit loss due to the differences in outcomes and inputs used for each of these (EBA, 2016, pp. 8).

The Basel Committee on Banking Supervision (BCBS) in December 2015 issued the document "Guidance on accounting for expected credit losses" (BCBS, 2015a) which thoroughly explains the supervisory expectation for credit institutions relating to sound credit risk practices. The supervisory guidance on credit risk and accounting for expected credit losses sent out by BCBS is a sound credit risk practice that will greatly benefit credit institutions with the implementation and application of the expected credit loss accounting model (EBA, 2016, pp. 4).

This thesis will also focus on developing a retail bank LGD model for expected credit loss to be used in the impairment calculation as set out in the IFRS 9 standard. This will be achieved by adapting a Basel LGD model.

1.3. Motivation

In this thesis, we study the estimation of the LGD component for both Basel and IFRS 9. Retail credit products can be classified into secured and unsecured products, and two separate approaches are described in this thesis to predict LGD. The difference between a secured and unsecured loan is the presence or absence of collateral. Collateral is given as security for possible non-repayment of a loan. A direct modeling approach is used to model LGD for unsecured products and an indirect modeling approach is used to model the LGD for secured products.

A direct modelling approach is followed by Witzany, Rychnovsky and Charamza (2012), which produces an EAD weighted LGD using survival analysis (denoted by EWSA), but Basel requires the LGD estimate to be default weighted. This can be seen in Paragraph 468 of the Basel Accord (BCBS, 2006) which states that: "This LGD cannot be less than the long-run default-weighted average loss rate given default calculated based on the average economic loss of all observed defaults within the data source for that type of facility. In addition, a bank must consider the potential for the LGD of the facility to be higher than the default-weighted average during a period when credit losses are substantially higher than average". In Chapter 2 we will enhance the EWSA approach by Witzany et al. (2012). The enhanced default weighted LGD model using survival analysis (denoted by DWSA) will produce default weighted LGD estimates, incorporating negative cashflows and will cater for over recoveries. The DWSA aligns more closely to regulations, since the Basel Accord requires default weighting. The model aligns more closely to actual data, since the model allows for over and negative recoveries that occur on LGD databases. It is expected that the predictive power of the model should improve. This assumption is tested on three retail bank datasets and five simulated datasets.

The indirect modelling approach followed by Leow and Mues (2012) modelled LGD by modelling a probability component, and loss severity component and combining them to estimate LGD. Incomplete accounts are excluded from the development of the probability component, since a binary target with outcomes write-off and not write-off is used. Valuable information is contained in incomplete accounts (EBA, 2016, pp. 34). LGD will be underestimated when incomplete accounts are excluded from LGD model development estimates. In Chapter 3, survival analysis is proposed as an improvement of the logistic regression approach used to model the probability component.

The adapted methodology also allows for censoring and the inclusion of incomplete accounts into the model development. The MSE, bias and variance of the two approaches are compared and it is shown that the change improves the predictive power of the model.

IFRS 9 replaced IAS 39 and moves from estimating actual losses to expected losses. Basel model development methodologies naturally lend itself to estimating the IFRS 9 LGD. The direct Basel LGD model development methodology of Chapter 2 will be adapted to model the IFRS 9 LGD in Chapter 4. The default date and months since default is used as segmentation in the DWSA model. A change for this segmentation scheme is required when adapting the model for IFRS 9. The lifetime of the account forms the basis of the expected credit loss calculation, and the month on book and application date characteristics, are used as segmentation for IFRS 9. The IFRS 9 LGD models are calibrated to recent information (Chawla, Forest and Aguais, 2016). LGD is modelled by month on book and a separate survival curve is created for every month on book. The DWSA methodology is used to create each of these survival curves.

2. Thesis objectives

A stable financial system is essential for growth in banks. A financial crisis can damage banks, as was seen in the financial crisis of 2008. Banks are regulated to reduce the risk of future financial crises. The accuracy of risk drivers to predict capital and impairments under Basel and IFRS 9, respectively, are key to the stability of banks. LGD is a key risk driver when estimating expected and unexpected losses. Given the key importance of the LGD for banks, the objectives of this thesis is to develop LGD models for Basel and for IFRS 9 that adhere to the required regulations and that are accurate.

LGD model developments need to align to the Basel Accord to give comfort to regulators and clients that the level of regulatory capital kept by a bank is sufficient to cover any possible unexpected losses. The direct LGD model development approach followed by Witzany et al. (2012) produces an EAD weighted LGD (denoted by EWSA). The **first objective** is to adopt the EWSA to produce a default weighted LGD (denoted by DWSA). Further enhancements made to the methodology are the inclusion of negative recoveries and incorporating over recoveries.

The indirect modelling approach followed by Leow and Mues (2012) models the probability and the severity component separately for LGD. The probability component is modelled by making use of logistic regression and a binary outcome. Logistic regression will be replaced by survival analysis and incomplete accounts will be modelled as a separate outcome, deriving at the **second objective**. The accuracy of Basel LGD models will increase by incorporating the information within incomplete accounts into the LGD modelling.

The **third objective** is to develop a robust IFRS 9 LGD modelling methodology. This methodology will be adapted from the direct Basel LGD model development methodology. Forward-looking LGD values will be predicted and adapted for macro-economic scenarios. The forward-looking macro-economic adjusted LGD values will be combined with a marginal PD and an EAD value to calculate expected credit losses on a portfolio.

3. Literature summary

This section will provide a short literature summary on unexpected credit loss under Basel and the expected credit loss under IFRS 9. LGD is a component of both expected and unexpected credit loss. The next subsection will provide a brief summary of LGD model methodologies. Lastly we will provide an introduction to survival analysis.

3.1. Expected and unexpected loss

The expected loss is the result of doing business, and banks typically manage these losses through pricing and provisioning. Losses over and above the expected losses are unexpected losses and, the bank needs to hold a buffer of capital against these losses. Capital is held to ensure that the bank meets regulatory obligations to cover unexpected losses where provisions is held to cover expected loss (BCBS, July 2005, pp. 2).

The risky position of a bank can be expressed in terms of a loss distribution. Typically, the expected loss is the expected value of the loss distribution. The unexpected loss may be defined as some risk measure of the loss distribution (e.g. value-at-risk), see Figure 1, or the difference between such a risk measure and the expected loss.

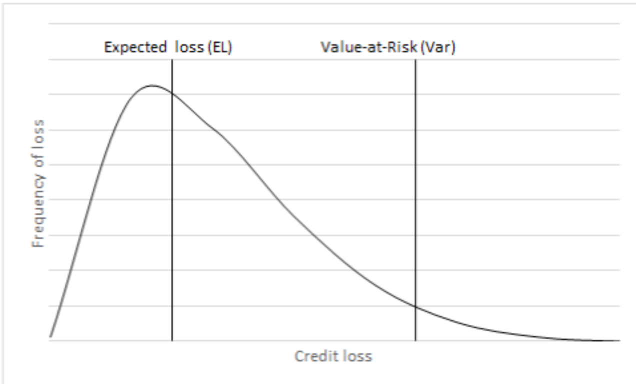


Figure 1: Loss distribution

3.1.1. Unexpected losses under Basel II

The loss probability density function is used by Basel to derive the capital formula. Figure 1 gives an example of a typical loss distribution. The right-skewed distribution shows that smaller losses occur more frequently. A confidence level is set equal to the likelihood that the bank remains solvent; the quantile of one minus this confidence level is equal to the value-at-risk. Under Basel, the unexpected loss is defined as the difference between the value-at-risk and the expected loss.

In the Basel II Accord (BCBS, 2006, pp. 52), banks adopting the advanced Internal Rating-Based (IRB) approach are allowed to model their own estimates for regulatory capital. The risk components that make up regulatory capital include measures of the PD, LGD and EAD. For the purpose of this thesis, the LGD component will be analysed and expanded. The risk weighted asset formula (RWA) for a retail portfolio in the capital Accord is

$$\text{RWA} = 12.5 \times \text{EAD} \times \text{LGD} \times (\Phi\left(\frac{\Phi^{-1}(\text{PD}) - \Phi^{-1}(0.999)\sqrt{\rho}}{\sqrt{1-\rho}}\right) - \text{PD}).$$

The correlation, ρ, measures the bank's exposure to the general state of the economy and Φ indicates the standard normal distribution.

LGD is the economic loss incurred by the bank when a customer defaults on a loan and is expressed as a fraction of EAD that is unpaid (BCBS, 2005, pp. 61). There exists a direct relation between LGD and the required capital that needs to be maintained. A 10% error in LGD will translate into a 10% error in regulatory capital. Due to the sensitivity of the regulatory capital formulae to LGD, it is necessary to ensure that the LGD estimation process is as accurate as possible (Witzany et al., 2012, pp. 20).

The Basel Accord states that the long run default weighted average LGD must be used and LGD estimates may not be lower than this value. Therefore the LGD is measured over a period that reflects economic downturn conditions. The data used in the development of an LGD model must span an economic cycle and needs to be at least five years for retail exposure and seven years for corporate exposures. The point in time LGD will vary with the economic cycle and the downturn LGD is therefore used for regulatory capital calculation (Engelmann and Rauhmeier, 2011, pp. 153).

3.1.2. Expected losses under IFRS 9

The expected credit loss (ECL) for account i that is currently at month on book m is:

$$ECL_{i,m} = (1+e)^{-h} \sum_{h=0}^{H} PD_{i,m,m+h} LGD_{i,m+h} EAD_{i,m+h}$$

The marginal $PD_{i,m,m+h}$ is the probability of account i defaulting at month on book $m+h$, given that the account remained preforming until month on book m. $LGD_{i,m+h}$ is the loss given that account i defaulted at month on book $m+h$ and $EAD_{i,m+h}$ is the exposure of account i that defaulted at month on book $m+h$. The cashflows on accounts are discounted to the reporting date by applying the current monthly effective interest rate (e).

The IFRS 9 standard requires the ECL estimates to be forward-looking and adjusted for macroeconomic scenarios. The time horizon, H, for the forward-looking information will vary between 12 months and remaining lifetime depending on the stage that the account is in. A stage is assigned based on changes in credit quality since initial recognition. Stage 1 is assigned when credit risk has not increased significantly since initial recognition. Stage 2 is assigned when credit risk has increased significantly since initial recognition. Stage 3 is assigned when an account defaults. A 12-month ECL is recognized for Stage 1 accounts and a lifetime expected loss (EL) is recognized for Stage 2 and Stage 3 accounts.

3.1.3. Basel vs IFRS 9

Now that we have discussed unexpected losses under Basel II and expected losses under IFRS 9, we will highlight the main differences between Basel and IFRS 9:

- Unexpected losses are calculated under Basel and used to predict the amount of capital that a bank needs to hold. The impairment calculation in a bank is regulated by the IFRS 9 standard and the expected credit losses are used to determine the provision used for impairments.
- The Basel models are predicted over a downturn period while IFRS 9 models make use of recent information.
- Indirect expenses are not added to the LGD for IFRS 9, but is added for the Basel LGD.
- A single default definition is used for IFRS 9 and a multiple default definition is used when capital is calculated as per the Basel Accord.
- Forward-looking information and macro-economic scenarios are used for IFRS 9 and not for Basel.

It is noteworthy that whilst Basel regulations are focussing on simplifying its approach, the IFRS 9 is getting more complicated in their respective approaches (moving from an incurred to an expected loss methodology) for calculating expected losses. It is of the utmost importance that the two regulations governing accounting and prudential standards are consistent with each other. A lack of consistency between the two mentioned pieces of legislation leads to complications and uncertainty for banks and regulators alike (De Jongh, Verster, Reynolds, Joubert and Raubenheimer, 2017, pp. 270–271).

Literature summaries on LGD methodologies and more specifically on workout LGD methodologies to directly and indirectly model the LGD are given in the next subsections.

3.2. Literature summary on LGD methodologies

A distinction can be made between subjective and objective LGD methodology. Subjective LGD methodology makes use of expert judgement and are used for low default portfolios, portfolios with insufficient data and new portfolios. Objective LGD methodology can be classified into the explicit and implicit methodologies. The explicit methodology allows for the direct computation of LGD, whereas with implicit methodology LGD relevant information needs to be extracted by applying applicable procedures. The market LGD, implied market LGD and the workout LGD are categorized as objective LGD methodology and expert judgement is categorized as a subjective method (Engelmann and Rauhmeier, 2011, pp. 157). The workout LGD is used in the retail sector, and the market LGD and implied market LGD are applied to the corporate sector. The market LGD is calculated as one minus the recovery percentage derived from the corporate bond price or share price available at the point of default. The implied market LGD is modelled from risky but not defaulted corporate bond or share prices by making use of a theoretical asset pricing model (BCBS, 2005, pp. 4).

The workout LGD (the focus of this thesis) can be modelled by using the direct approach or the indirect approach. When using the direct approach, the LGD is equal to one minus the recovery rate (De Jongh et al., 2017, pp. 261). The indirect approach uses two components that are modelled separately, namely the probability component and the loss severity component. The market LGD is an example of an ex-post or actual LGD and the workout LGD is an example of an ex-ante or estimated LGD (Engelmann and Rauhmeier, 2011, pp. 157 – 158). Figure 2 contains a diagram that illustrates the classification of the various LGD approaches.

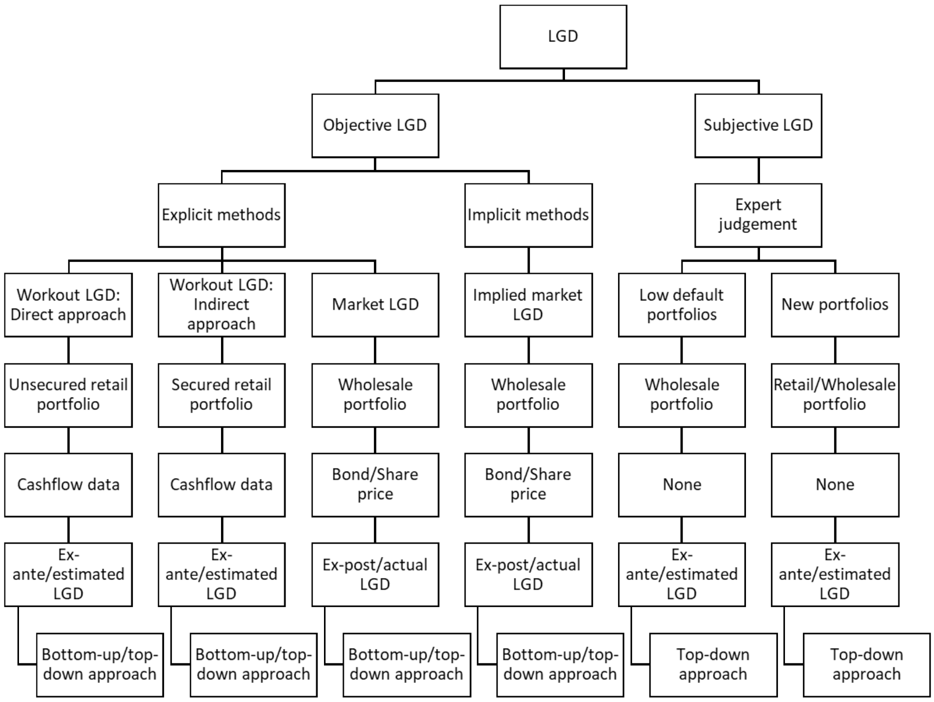


Figure 2: LGD approaches classified

The workout LGD models used in the retail sector are not as advanced as the market LGD or implied market LGD models used in corporate loans due to the fact that most of the work on prediction of LGD pertains to the corporate sector (Qi and Yang, 2009, pp. 788). Corporate bond prices and share prices are publicly available at the point of default and used to infer the relative credit risk of the underlying company, the associated risk premium and the recovery percentage (Leow and Mues, 2012, pp. 184). The papers by (Lotheram, Brown, Martens, Mues and Baesens, 2012), (Qi and Zhao, 2011) and (Bellotti and Crook, 2012) contain comparisons between different LGD modelling techniques. This thesis will focus on workout LGD in the retail sector.

3.3. Literature summary on workout LGD to directly model LGD

The workout LGD is equal to one minus the recovery rate, where the recovery rate can be calculated as the sum of all future recoveries discounted to the default point expressed as a percentage of the exposure at default.

Witzany et al. (2012) propose a direct modelling approach using EAD weighted survival analysis using a Cox proportional hazard model. Other methods used in literature to model LGD include beta regression, ordinary least squares, fractional response regression, inverse beta transformation, run-off triangle and Box-Cox transformation.

Although a run-off triangle (Braun, 2004, pp. 401) is most often used, it cannot take covariates into account. A separate run-off triangle needs to be created for every segment or attribute of a covariate. In the other mentioned methods, the covariates can be grouped into attributes and modelled onto the LGD. In the beta regression suggested by Brown (2014, pp. 65 – 66), a beta distribution is fitted to the LGD. The beta distribution is reparametrized and covariates are modelled onto the new parameters. For the ordinary least squared approach, a linear regression is used to model LGD directly (Witzany et al., 2012, pp. 12). The LGD is the dependant variable in the linear regression and the covariates are modelled onto LGD.

Bastos (2010, pp. 2512) describes the fractional response regression. The LGD is taken as the dependant variable. The Bernoulli log likelihood is maximized to estimate the parameters. A logistic function is used for the functional form. Brown (2014, pp. 64) described the inverse beta model in his article. He applies a cumulative beta distribution to the recovery rate and estimates the parameters. The inverse standard normal cumulative distribution function is then applied in reverse to get the predicted LGD.

Braun (2004, pp. 401) describes the run-off triangle approach. Recovery amounts are summed by default date and months since default. The available recovery information forms a triangle. The available recovery information is used to predict future recovery information by applying a technique called the chain ladder approach. The Box-Cox transformation is applied to the recovery rate variable. An ordinary least square is applied to the transformed variable and the transformation is applied in reverse (Brown, 2014, pp. 66).

A more detailed description of these various LGD modelling methodologies from literature are given in Appendix A of the article titled “Default weighted survival analysis to directly model loss given default” (Chapter 2).

3.4. Literature summary on workout LGD to indirectly model LGD

The indirect approach uses two components that are modelled separately, namely the probability component and the loss severity component. Somers and Whittaker (2007) introduced the idea of an indirect LGD model whereby the LGD is calculated by combining a probability component and a haircut (loss severity) component, but did not detail the development of the probability component. The paper by Leow and Mues (2012, pp. 183) describes the indirect approach whereby LGD is calculated by combining two models. The two models are the haircut model and the probability model. The probability model provides an estimate of the probability of each account undergoing a loss event. The haircut model predicts the difference between the forced sale price and the market valuation of the repossessed property (Leow and Mues, 2012, pp. 186).

Although there are limited literature on indirect LGD methodologies, there exist a vast number of modelling techniques on modelling the probability of an event and the severity of an event. However, this is outside the scope of this thesis.

3.5. Survival analysis

The set of procedures used to study data, which concludes in a specific event (such as death or inability to repay a loan), is called survival analysis. To measure and study the time up to the occurrence of that specific event is the main aim of survival analysis. Survival analysis is used across this paper to model loss given default and therefore a literature study of survival analysis is contained in this section.

The survival and hazard functions are set out in the section below. The hazard- and survival functions is relevant for survival analysis, and therefore a literature review on this topic follows.

The influence of covariates on the hazard function is modelled through the proportional hazard model. In practice, we use the proportional hazard model to show how covariates influence the LGD, hence the relevance in reviewing the concept in this literature study. The proportional hazard model makes use of assumptions that two events cannot occur simultaneously, however in practice, two deaths can occur at the same time, similar to two defaults coinciding. We therefore need to make provision for such occurrences, and the treatment of ties is reviewed in the following section. The parameter estimates for the Cox-proportional hazards model are fitted by making use of the Newton-Raphson procedure, as is set out below.

Positive and negative survival curves will be combined throughout this thesis, in an LGD model development section, and an empirical proof is therefore given here through making use of an example.

3.5.1. The Survival and hazard function

According to Collett (2003, pp. 11-13), when reviewing survival data, there are two functions that are relevant: the survival function and the hazard function. The remaining life span of a human, t, can be expressed as a variable T. The various T values can be expressed as a probability distribution. Survival time is assigned the random variable of T. The probability density function underlying to the probability distribution of T is described as $f(t)$. The distribution function of T is then stated as

$$F(t) = P(T < t) = \int_0^t f(u)du.$$

This represents the likelihood that the survival time is less than random variable, t. A survival function $S(t)$ is then set as

$$S(t) = P(T \geq t) = 1 - F(t).$$

This is the likelihood that survival time is bigger than or equal to t. A survival function can be used to determine the probability of a person surviving from inception to after point t. The hazard function is used to express the risk of death at point t. The risk of death is calculated based on the probability of an individual surviving past point t.

An individual's survival time T lies somewhere between t and $[t+dt]$. T is greater or equal to t, denoted as $P(t{\leq}T{<}t{+}dt|T{\geq}t)$. The probability conditional upon the aforementioned, is then converted to a rate by expressing it as a probability per time unit (by dividing by the time interval). Finally, the hazard function is written as the limiting value of the aforementioned quantity as dt approaches zero,

$$h(t) = \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}.$$

From the above, $h(t)dt$ is therefore an estimation of the likelihood that an individual dies within the time interval $(t, t+dt)$, qualified by that person surviving for time t. Simplistically, the hazard function describes the risk of dying at point t.

There are some relationships of interest between the hazard and survival calculations: The likelihood of event A, provided the occurrence of event B, is given by $P(A|B)$ $=P(AB)/P(B)$, where

$P(AB)$ is the likelihood of both A and B occurring. By incorporating the result of the aforementioned, the conditional probability of the hazard function is denoted as

$$\frac{P(t \leq T \leq t + \Delta t)}{P(T \geq t)},$$

which is equal to $\frac{F(t+\Delta t)-F(t)}{S(t)}$, where F(t) is the distribution function of T. Then,

$$h(t) = \frac{F(t+\Delta t) - F(t)}{\Delta t S(t)}.$$

The hazard function is then equal to

$$h(t) = \frac{f(t)}{S(t)}.$$

It then follows that

$$h(t) = -\frac{d}{dt}\log(S(t)),$$

and so

$$S(t) = \exp(-H(t)),$$

where

$$H(t) = \int_0^t h(u)du.$$

The cumulative hazard can be obtained from the survivor function, since

$$H(t) = -\log(S(t)).$$

3.5.2. The proportional hazard model

The effect of covariates on the hazard rate of an individual is modelled through the proportional hazard model.

Collett (2003, pp. 63-64) gives the proportional hazard model, for the i th individual with p explanatory variables, as

$$h_i(t) = h_0(t)\, e^{\beta_1 x_{1i} + \ldots + \beta_p x_{pi}}.$$

There are two components to this equation, the coefficients, x_i, in the linear component of the model and the baseline hazard function, $h_0(t)$. These two components can be obtained separately. First the beta values, β_i, are approximated and these estimations are then used to approximate the baseline hazard function. This has the effect that we do no require an estimate of $h_0(t)$ in order to deduce the effect of the explanatory variables on the relative hazard, $\frac{h_i(t)}{h_0(t)}$.

Assume that there are n individuals with r deaths occurring. There are n minus r survivals, which are censored. For simplicity, the assumption is introduced that there are no ties, in that only one death occurs at a time. Death times will be ordered up to r and denoted by $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Time $t_{(j)}$ will be the jth ordered death. Individuals that are at risk of death at $t_{(j)}$ are written as $R(t_{(j)})$. This is the group of individuals that are uncensored and alive just before $t_{(j)}$. The quantity so denoted is called the risk set.

According to Cox (1972) the likelihood function for the proportional hazards model given above as is given as

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' \mathbf{x}_{(j)})}{\sum\limits_{l \in R_{(t_{(j)})}} \exp(\beta' \mathbf{x}_l)},$$

where $\mathbf{x}_{(j)}$ is the vector of covariates for the person that dies at the jth ordered death time $t_{(j)}$. The sum in the denominator of the function is the sum of the values of $\exp(\beta' \mathbf{x}_{(j)})$ of all persons who are at risk of dying at time $t_{(j)}$. The product is across all persons for whom a recorded death time exists. Censored persons do not feature in the numerator of the function, but feature in the calculation in the summation of the risk sets at death time, where it occurs before the censored time. The ranking of the time of death determines the likelihood function, as this points to the risk set at each death time. Synopses about the impact that explanatory variables have on the hazard function are only determined by the rank order of the survival times.

If the assumption is made that the data is comprised of n observed survival times, $t_1, \ldots, t_n$, and where δ_i is an indication of an event, which assumes the value of zero or one. δ_i takes a value of one, where the ith survival time is not right censored, and zero where it is. The likelihood function can then be expressed as

$$\prod_{i=1}^{n} \left(\frac{\exp(\beta' \mathbf{x}_{(i)})}{\sum\limits_{l \in R_{(t_i)}} \exp(\beta' \mathbf{x}_l)} \right)^{\delta_i},$$

where $R(t_i)$ is the risk set at time t_i. The corresponding log-likelihood function is then expressed as

$$logL(\beta) = \sum_{i=1}^{n} \delta_i \{\beta' \mathbf{x}_i - \log \sum_{l \in R_{(t_i)}} \exp(\beta' \mathbf{x}_l)\}$$

By maximizing this log-likelihood function with the numerical technique, the maximum likelihood estimates of the beta values in the proportional hazards method can be found.

3.5.3. Treatment of ties

Collett (2003, pp. 67) assumed that the hazard function is continuous and that simultaneous (tied) survival times are impossible. In practice, survival times are often rounded to the nearest day, month or even year.

There can be multiple deaths at the same time, or more than one censored observation at death time. Where the aforementioned occur simultaneously, we introduce the assumption that censoring takes place after the deaths. The conundrum of which deaths should be included in the risk set at the time of death, is removed and tied, censored observations present no further problems in the likelihood function. Tied survival times need only now be considered in fitting the proportional hazards model.

The likelihood function now needs to be modified to provide for tied observations. Let $\mathbf{s}_j$, contain the sum of p covariates who died at the jth death time, $t_j, j = 1 \ldots r$. If there are d_j deaths at $t_{(j)}$ the hth element of $\mathbf{s}_j$ is $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$ where x_{hjk} is the value of the h th explanatory variable, $h = 1, .., p$

for the kth of d_j individuals, $k = 1, \ldots d_j$ who died at the jth death time, $j = 1, \ldots, r$. The Breslow likelihood approximation is then

$$\prod_{j=1}^{r} \frac{\exp(\beta' \mathbf{s}_j)}{\left(\sum_{l \in R_{(t_{(j)})}} \exp(\beta' \mathbf{x}_l)\right)^{d_j}}.$$

Efron proposed the following approximate likelihood function

$$\prod_{j=1}^{r} \frac{\exp(\beta' \mathbf{s}_j)}{\prod_{k=1}^{d_j} [\sum_{l \in R_{\left(t_{(j)}\right)}} \exp(\beta' \mathbf{x}_l) - (k-1) d_j^{-1} \sum_{l \in D_{(t_{(j)})}} \exp(\beta' \mathbf{x}_l)]},$$

where $D_{(t_{(j)})}$ is the set of individuals who died at $t_{(j)}$. Both the Breslow and Efron approximation are used in practice and give similar results.

3.5.4. The Newton-Raphson procedure

Cencored survival analysis models are fitted using the Newton-Raphson procedure to maximize the partial likelihood function. A description of this procedure is set out below.

The beta values, at the $(s+1)$th cycle, can be estimated by applying the iterative procedure

$$\hat{\beta}_{s+1} = \hat{\beta}_s + \mathbf{I}^{-1}(\hat{\beta}_s)\mathbf{u}(\hat{\beta}_s).$$

The $\mathbf{u}(\hat{\beta}_s)$ is the first derivative of the log-likelihood function with respect to $\hat{\beta}_s$ and $\mathbf{I}(\hat{\beta}_s)$ is the second derivative of the of the log-likelihood with the (jk)th element of $\mathbf{I}\left(\hat{\beta}_s\right)$ equal to

$$-\frac{d^2 logL\left(\hat{\beta}_s\right)}{d\hat{\beta}_j d\hat{\beta}_k}.$$

The inverse of this information matrix is taken and used in the iterative equation. An initial value of $\hat{\beta}_0 = 0$ can be taken. The process can be repeated until the change in parameter estimates are negligibly small.

3.5.5. Combining positive and negative survival curves

Different methods to combine survival curves exist, in the section a illustrative example of such a possibility is discussed. The above literature study describes survival analyse in general. The below example is specific to Loss given default model development.

The purpose of the following example is to illustrate that

$$S(t) = S_p(t) + (1 - S_n(t)).$$

The values for the combined survival curve,$S(t)$, is calculated in Table 1. The survival curve, $S(t)$ is calculated as the sum of the exposure at default (EAD) values, less the sum of the cashflows up to point t in default, divided by the EAD.

t	EAD	1	2	3
Account A Cashflow	100	20	-30	60
Account B Cashflow	250	150	320	-10
Account C Cashflow	320	180	10	18
Total	670	350	300	68
$S(t)$	100.00%	47.76%	2.99%	-7.16%

Table 1: Overall survival curve

The values for $S(t)$ in Table 1 is calculated as follows.

$$S(1) = \frac{670 - 350}{670} = 47.76\%$$

$$S(2) = \frac{670 - 350 - 300}{670} = 2.99\%$$

$$S(3) = \frac{670 - 350 - 300 - 68}{670} = -7.16\%$$

The negative cashflow values were excluded from Table 1 and only positive cashflows were included in Table 2. The values for the positive survival curve is calculated in Table 2.

t	EAD	1	2	3
Account A Cashflow	100	20		60
Account B Cashflow	250	150	320	
Account C Cashflow	320	180	10	18
Total	670	350	330	78
$S_p(t)$	100.00%	47.76%	-1.49%	-13.13%

Table 2: Positive survival curve

Here is the details to calculate $Sp(t)$ in Table 2.

$$S_p(1) = \frac{670 - 350}{670} = 47.76\%$$

$$S_p(2) = \frac{670 - 350 - 330}{670} = -1.49\%$$

$$S_p(3) = \frac{670 - 350 - 330 - 78}{670} = -13.13\%$$

The negative chashflows are kept in Table 3 and the negative survival curve, $S_n(t)$, calculated.

The negative survival curve values, $S_n(t)$, is calculated as follow.

$$S_n(1) = \frac{670}{670} = 100\%$$

t	EAD	1	2	3
Account A Cashflow	100		30	
Account B Cashflow	250			10
Account C Cashflow	320			
Total	670	0	30	10
$S_n(t)$	100.00%	100.00%	95.52%	94.03%

Table 3: Negative survival curve

$$S_n(2) = \frac{670-30}{670} = 95.52\%$$
$$S_n(3) = \frac{670-30-10}{670} = 94.03\%$$

The positive and negative survival curves from Table 2 and Table 3 are given below and the value $S_p(t) + (1-S_n(t))$ calculated.

t	0	1	2	3
$S_p(t)$	100.00%	47.76%	-1.49%	-13.13%
$S_n(t)$	100.00%	100.00%	95.52%	94.03%
$S_p(t) + (1-S_n(t))$	100.00%	47.76%	2.99%	-7.16%

Table 4: Combining positive and negative survival curves

The value of $S_p(t) + (1-S_n(t))$, from Table 4, is equal to $S(t)$, from Table 1. This example gives empirical proof that $S(t) = S_p(t) + (1-S_n(t))$.

4. Summary

Retail banks use the Basel LGD as one of the estimates to calculate regulatory capital and this forms the focus of Chapter 2. This chapter describes the Basel LGD for the direct approach. The Basel LGD is modelled directly by estimating the LGD as one minus the recovery rate. The basis for Chapter 2 is based on Witzany et al. (2012) who follow a direct modelling approach. Various LGD modelling methodologies are compared in a simulation study and on retail data.

While Chapter 2 describes the direct approach, Chapter 3 describes the indirect approach. The indirect approach uses two components that are modelled separately, being the probability component and the loss severity component. The indirect approach is applied in a simulation study using selected parameters to give similar survival curves as that of a retail bank's vehicle and asset portfolio and home loans portfolio. The probability component of the Basel LGD is determined by using survival analysis and logistic regression approach respectively. These two approaches are applied to simulated datasets and the mean squared error (MSE), bias and variance are compared.

Where the focus of Chapters 2 and 3 were on LGD models for regulatory capital to cover unexpected losses, the focus in Chapter 4 is shifted to LGD models for provisions to cover expected losses. The Basel III (BCBS, July 2010) Accord also supports the move from the incurred losses provisioning approach to the expected loss provisioning approach. This motivates the adjustment of the Basel LGD models to IFRS 9 LGD models in Chapter 4.

The IFRS 9 LGD is used to predict the provisioning that is needed to cover the expected credit losses depicted in IFRS 9 accounting standard. The IFRS 9 LGD is calculated for every age an account can reach, where lifetime refers to the maximum age an account will reach. Chapter 4 introduces a new model for the IFRS 9 LGD where credit losses are calculated by using a forward-looking lifetime LGD and an adjustment for macro-economics. Given that banks only recently adopted the IFRS 9 standard (from January 2018), limited literature is available with regard to this topic.

Chapter 5 concludes the thesis. The key findings of the thesis are summarised and further research ideas are provided.

References

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2005). Studies on the validation of international rating systems. Working paper 14.

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2006). International convergence of capital measurement and capital standards. *URL:https://www.bis.org/publ/bcbs128.pdf.*

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2014). Revisions to the standardized approach for credit risk.

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2015a). Guidance on accounting for expected credit losses. *URL:https://www.bis.org/bcbs/publ/d311.pdf/.*

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2015b). A brief history of the Basel committee.

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (July 2005). An explanatory note on the Basel II IRB risk weight functions.

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (July 2010). Basel iii: A global regulatory framework for more resilient banks and banking systems.

BASTOS, J. (2010). Forecasting bank loans loss given default. *Journal of banking and Finance, 34. 2510-2517.*

BELLOTTI, T. AND CROOK, J. (2012). Loss given default models incorporating macro-economic variables for credit cards. *International Journal of Forecasting 28 (2012) 171182.*

BRAUN, C. (2004). The prediction error of the chain ladder method applied to correlated run-off triangles. *Astin Bulletin, Vol. 34, No. 2, 2004, pp. 399-423.*

BROWN, I. (2014). Developing credit risk models using SAS enterprise miner and sas/stat: Theory and application. *Cary, NC:SAS Institute inc.*

CHAWLA, G., FOREST, L., AND AGUAIS, S. (2016). Point-in-time (pit) LGD and EAD models for IFRS9, cecl and stress testing. *URL:http://www.henrystewartpublications.com/jrm/.*

COLLETT, D. (2003). Modelling survival data in medical research. Chapman and Hall.

DEJONGH, P., VERSTER, T., REYNOLDS, E., JOUBERT, M., AND RAUBENHEIMER, H. (2017). A critical review of the Basel margin of conservatism requirement in a retail credit context. *Inter- national Business and Economics Research Journal Fourth Quarter 2017 Volume 16, Number 4.*

ENGELMANN, B. AND RAUHMEIER, R. (2011). The Basel II risk parameters. estimation, validation, stress testing with applications to loan risk management. *Springer Heidelberg Dordrecht London New York.*

EUROPEAN BANKING AUTHORITY (EBA) (2016). Consultation paper EBA/cp/2016/10: Draft guidelines on credit institutions: credit risk management practices and accounting for expected credit losses. *URL: https://www.eba.europa.eu/documents/10180/1532063/EBA-CP-2016-10+%28CP+on+Guidelines+on+Accounting+for+Expected+Credit%29.pdf.*

GLOBAL PUBLIC POLICY COMMITTEE (GPPC) (2016). The implementation of IFRS 9 impairment requirements by banks: Considerations for those charged with governance of systemically important banks. *URL:http://www.ey.com/Publication/vwLUAssets/Implementation_of_IFRS_9_impairment_requirements_by_systemically_important_banks/$File/BCM-FIImpair-GPPC-June2016%20int.pdf.*

IFRS (2014). IFRS9 financial instruments: Project summary. *URL:http://www.ifrs.org/Current-Projects/IASB-Projects/Financial-Instruments-A-Replacement-of-IAS-39-Financial-Instruments-Recognitio/Documents/IFRS-9-Project-Summary-July-2014.pdf*.

LEOW, M. AND MUES, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting, pp 183195.*

LOTHERAM, G., BROWN, I., MARTENS, D., MUES, C., AND BAESENS, B. (2012). Benchmarking regression algorithms for loss given default modelling. *International Journal of Forecasting 28 (2012) 161170*.

QI, M. AND YANG, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance 33 (2009) 788799*.

QI, M. AND ZHAO, X. (2011). Comparison of modelling methods for loss given default. *Journal of Banking and Finance 35 (2011) 28422855*.

SOMERS, M. AND WHITTAKER, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research, 183, 14771487*.

WITZANY, J., RYCHNOVSKY, M., AND CHARAMZA, P. (2012). Survival analysis in LGD modelling. *European Financial and Accounting Journal, 2012, vol. 7, no. 1, pp. 6-27*.

Chapter 2

Default Weighted Survival Analysis to Directly model Loss Given Default.

Chapter 2
Section 1

Guidelines for authors submitting an article to the South African Statistical Journal.

South African Statistical Journal
Suid-Afrikaanse Statistiese Tydskrif

Official format of the South African Statistical Journal

All articles published in the South African Statistical Journal should adhere to the following set of guidelines to ensure uniformity and consistency of publications. An example can be obtained from the managing editor on request (leonard.santana@nwu.ac.za). It is preferable that articles are submitted using the LaTeX PDF format (the LaTeX template can be found at http://sastat.org.za/sites/default/files/files/SASJ%20PDF%20LaTeX%20template(2).zip, but, for the initial phase of screening, MS Office Word or Scientific Word documents are also accepted.

Biographical information

The biographical information should contain the name of all authors in the form initials then surname, e.g. U. N. Named. After each author's name comes the name of the institution of affiliation and a postal and / or e-mail address. In the case of multiple authors the corresponding author should be indicated in a footnote.

Key words

The key words should be listed above the abstract and should appear in alphabetical order.

For guidelines on the choice and importance of key words see Gbur and Trumbo (1995). The full reference is provided as an example in the references section.

Abstract

The aim of the abstract is to provide a concise description of your article. It should be no more than 250 words and contains a minimum of symbols and references.

Subject classification

This Journal uses the Mathematical Subject Classification 2000 (MSC2000) system. More information on the system can be found at www.ams.org/msc.

Main body

For the main body of the article the following guidelines should be adhered to:

- Sections are numbered consecutively using Arabic numerals.
- The first paragraph of each section or subsection has no indentation of the left margin.
- All subsequent paragraphs in a section or subsection are indented at the left margin.
- 1.2 spacing is used except for list items.
- Full stops are not used after theorems, remarks, lemmas, corollaries and examples (e.g. Theorem 1 or Example 1).
- Full stops are used after tables, figures and proofs (e.g. Proof. or Table 1.).
- Equations that referenced in the text must be numbered sequentially on the right hand side of the page using Arabic numerals. Equations that are not referenced should not receive an equation number.
- Displayed equations (i.e., equations that are appear on their own line of text and are centred on the page) should contain appropriate punctuation.
- Equations are referenced by simply stating the equation number in parentheses, e.g., "(1)" or "(4)". It is no necessary to use the word "Equation" when referencing, that is, do not write "Equation (1)."
- When references are cited as nouns, then they must be written either as "Abramowitz and Stegun (1970)" or as "Abramowitz and Stegun (1970, page 100)."
- Figures and tables may be submitted separately. Place one figure / table on a page and identify it clearly. Indicate the position in the text where you wish the figure / table to be placed with the (uppercase) phrase:

  INSERT FIGURE / TABLE X ABOUT HERE.

Bulleted or numbered items

Single spacing is used for both bulleted and numbered items. Sublevels of bullets and numbers are indented by the same width. The order of succession for lower level numbering is:

1. First level
    a. Second level
        i. Third level
            I. Fourth level

The order of succession for bullets is:

• First level
    ► Second level
        - Third level
            * Fourth level

Theorems, Lemmas and proofs

A theorem (lemma) is stated starting with the word Theorem (Lemma) in bold and numbered consecutively using Arabic numerals. The full stop is omitted. The theorem is then stated in the normal font. The proof starts with the word “Proof” in bold, then a full stop and the proof follows. The end of the proof is indicated by a solid square. An example:

Theorem 1 Here we state the theorem.
Proof. Here we prove the theorem. ■

Figures

Figure names and description are placed directly *below* the figure. Figures are numbered in the order that they are cited using Arabic numerals. The word “Figure” is in bold and the number is followed by a colon. The description follows in the normal font. If the description spans less than one line it is centred. Multiple lines are justified. An example:

Figure 1: Here we provide a short description.

Tables

Table names and descriptions are placed directly *above* the table. Tables are numbered in the order that they are cited using Arabic numerals. The word “Table” is in bold and the number is followed by a colon. Then a description follows in the normal font. If the description spans less than one line it is centred. Multiple lines are justified. An example:

Table 1: Here we provide a short description. If the description spans more than one line, it is justified.

The table follows.	

Appendices

Appendices are placed at the back of the article and numbered alphabetically if there is more than one, e.g. Appendix A: Descriptive title A, Appendix B: Descriptive title B, etc. No section number is used.

References

Only references cited in the text should be included. No section number is used. The format of references is illustrated by the following examples:

Book:

ABRAMOWITZ, M. AND STEGUN, I. (1970). *Handbook of Mathematical Functions.* Dover Publications: New York.

Article in a journal:

BOLLERSLEV, T., CHOU, R. Y., AND KRONER, K. F. (1992). ARCH modelling in Finance: a review of the theory and empirical evidence. *Journal of Econometrics*, **39**, 5–59.

GBUR, E. E. AND TRUMBO, B. E. (1995). Key words and phrases—The key to scholarly visibility and efficiency in an information explosion. *The American Statistician*, **49** (1), 29−33.

Proceedings article:

WOLFINGER, R. D. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. *In Proceedings of the 24th Annual SAS Users Group International Conference (SUGI 24)*. Miami Beach, FL, USA, pp. 278–284.

Chapter in a book:

BOLLERSLEV, T., ENGLE, R. F., AND NELSON, D. B. (1994). ARCH models. *In* ENGLE, R. F. AND MCFADDEN, D. C. (Editors) *Handbook of Econometrics*. North-Holland: Amsterdam, pp. 2959–3038.

Note the use of the “small caps” font for the author names. Note also that the titles of books have the first letter of each word capitalised, whereas the titles of journal articles employ normal sentence case (i.e., only the first letter of the first word and proper nouns are capitalised).

Acknowledgements

Acknowledgements may be included as a separate section before the references. Acknowledgements should be kept concise. No section number is used.

Chapter 2
Section 2

Article Title:

Default Weighted Survival Analysis to Directly model Loss Given Default.

Article Authors:

M. Joubert, T. Verster and H. Raubenheimer.

The article was published in the South African Statistical Journal 2018, Vol. 52, No. 2, 173-202.

Contents of Chapter 2 Section 2

List of Figures

List of Tables

South African Statistical Journal
2018, Vol. 52, No. 2, 173–202

DEFAULT WEIGHTED SURVIVAL ANALYSIS TO DIRECTLY MODEL LOSS GIVEN DEFAULT

Morne Joubert[1]
Centre for BMI, North-West University, Potchefstroom, South Africa
e-mail: *joubertmorne9@gmail.com*

Tanja Verster
Centre for BMI, North-West University, Potchefstroom, South Africa

Helgard Raubenheimer
Centre for BMI, North-West University, Potchefstroom, South Africa

Traditionally when predicting loss given default (LGD), the following models can be used: beta regression, inverse beta model, fractional response regression, ordinary least squares regression, survival analysis, run-off triangles and Box–Cox transformation. The run-off triangle method is commonly used in practice.

When using survival analysis to model LGD a standard method to use is exposure at default (EAD) weighted survival analysis (denoted by EWSA). This article will aim to enhance the survival analysis estimation of LGD. Firstly by using default weighted LGD estimates and incorporating negative cash flows and secondly catering for over-recoveries. We will denote this new method to predict LGD as the default weighted survival analysis (DWSA). These enhancements were motivated by the fact that the South African Reserve Bank requires banks to use default weight LGD estimates in regulatory capital calculations. Therefore by including this into the survival analysis approach, the model is aligned more closely to regulations. Recovery datasets used by banks include both negative and over-recoveries. By including these into the LGD estimation, the models more are closely aligned to the actual data. The assumption is that the predictive power of the model should therefore be improved by adding these changes. The proposed model is tested on eight datasets. Three of these are actual retail bank datasets and five are simulated. The datasets used are representative of the data typically used in LGD estimations in the South African retail environment.

This article will show that the proposed DWSA model outperforms the EWSA model by resulting in not only the lowest mean squared error (MSE), but also the lowest bias and variance across all eight datasets. Furthermore, the DWSA model outperforms all other models under review.

Key words: Basel, Direct modelling approach, Loss given default, Survival analysis.

1. Introduction and literature overview

Loss given default (LGD) is the loss incurred by a bank (economic loss) when a customer is unable to pay back a loan, and this is stated as the exposure at default (EAD) portion that remains unpaid.

LGD is one of the estimates that a retail bank uses to calculate regulatory capital and forms the focus of this article. LGD can either be modelled through the direct approach or the indirect

---

[1] Corresponding author.

MSC2010 subject classifications. 62-07, 62N01.

approach. The indirect LGD modelling approach combines two components namely the loss severity component and probability component. The probability component predicts the probability that a defaulted account will remain in default or that a loss will occur on this account. The loss severity component gives the value of the estimated loss. In this article, LGD will be modelled directly by estimating LGD as one minus the recovery rate. Witzany, Rychnovsky and Charamza (2012) propose a direct modelling approach using EAD weighted survival analysis (EWSA).

The Basel accord introduced the concept of long run default weighted average LGD. It is compulsory to use this measure (BCBS, 2006, par. 468). BCBS (2006) states that LGD estimates cannot materially differ from the long run default weighted average LGD. In the interest of aligning LGD estimates to the Basel accord this article proposes a default weighting survival analysis (DWSA) approach.

The DWSA approach further extends the EWSA approach by including negative cash flows into cash flow streams when modelling LGD and adapting methodology to cater for over-recoveries. Over-recoveries occur when more of the previously unpaid loan is received than the EAD. The proposed enhancements align the modelling approach to produce more accurate results and decrease the mean squared error (MSE) and bias of the model. Other methods used in literature to model LGD include beta regression, ordinary least squares, fractional response regression, inverse beta transformation, run-off triangle and Box–Cox transformation. We will compare our technique with these seven techniques, but first we will give a brief overview of each.

Although a run-off triangle (Braun, 2004, p. 401) is most often used, it cannot take covariates into account. A separate run-off triangle needs to be created for every segment or attribute of a covariate. In the other methods mentioned above, the covariates can be grouped into attributes and modelled onto the LGD. In the beta regression suggested by Brown (2014, pp. 65–66) a beta distribution is fitted to the LGD. The beta distribution is reparametrised and covariates are modelled onto the new parameters. For the ordinary least squares approach a linear regression is used to model LGD directly (Witzany et al., 2012, p. 12). The LGD is the dependent variable in the linear regression and the covariates are modelled onto LGD. Bastos (2010, p. 2512) describes the fractional response regression, where the LGD is taken as the dependent variable. The Bernoulli log-likelihood is maximised to estimate the parameters, and a logistic function is used for the functional form. Brown (2014, p. 64) describes the inverse beta model in his article, where he applies a cumulative beta distribution to the recovery rate and estimates the parameters. The inverse standard normal cumulative distribution function is then applied in reverse to get the predicted LGD. Braun (2004, p. 401) describes the run-off triangle approach. Recovery amounts are summed by default date and months since default. The available recovery information forms a triangle and is used to predict future recovery information by applying a technique called the chain ladder approach. The Box–Cox transformation is applied to the recovery rate variable. Ordinary least squares is applied to the transformed variable and the transformation is applied in reverse (Brown, 2014, p. 66).

Section 2 is dedicated to evaluating the contributions made by Witzany et al. (2012) (EWSA approach). Implementing the EWSA model is explored in this section. The new DWSA approach is discussed in Section 3, where the enhancements to the EWSA model are explained in detail. The three enhancements serve to incorporate common practice into modelling techniques and to align modelling of LGD to ruling legislation on this topic in the Basel accord (BCBS, 2006). Data specification and data simulation form the topic of Section 4. Actual retail bank data for credit cards, revolving loan and cheque accounts are used. Data simulation techniques that are representative of a typical retail bank's LGD modelling data are discussed. The fit of the model to the actual data is discussed and the MSE, bias and variance are described. Section 5 contains the results of both the retail bank data and simulated data. In both these sections the results are unanimous that the DWSA model outperforms all other models, including the EWSA model. The beta regression model is the second best performing model. The performance of these models are gauged by comparing the MSE, bias and variances. Section 6 concludes the article.

2. EAD weighted survival analysis (EWSA)

In this section, the EWSA approach by Witzany et al. (2012) is discussed. Witzany et al. (2012) use survival analysis to directly model LGD. The EWSA model is the starting point for the DWSA model discussed in the subsequent section.

2.1 Loss given default

Witzany et al. (2012, p. 8) distinguish between the market LGD and the workout LGD. The market LGD is calculated on instruments such as bonds and other debt instruments, while for other receivables a workout LGD is used. More specifically, the workout LGD is used when modelling unsecured retail credit portfolios (Witzany et al., 2012, p. 19). Market LGD is calculated as the market value over the face value shortly after the point of default.

The workout LGD assumes a workout process that ends T_w months after the default point and that no further recoveries are made past this point. Unsecured retail credit portfolios are used in this article and the workout LGD is used. Mathematically the workout LGD is expressed as

$$LGD_{i,0} = \frac{EAD_{i,0} - \sum_{t=1}^{T_w} DCF_{i,t}}{EAD_{i,0}},$$

where $DCF_{i,t} = CF_{i,t}/(1+r)^t$ is the discounted future cash flows for account i at time t. $LGD_{i,0}$ is the LGD value for account i at time $t = 0$. The recovery time t for a defaulted account i is measured in months and takes only values $\{1, 2,..., T_w\}$. Cash flows $CF_{i,t}$ are calculated as the difference between account balances now, versus account balances in the previous month, adding back the interest and the fees, subtracting the amount written off. The post write-off recoveries represent recovery or additional expense amounts post the write-off date, which are added to the cash flows (Witzany et al., 2012, p. 8). The rate r used to discount cash flows to the present value is represented by the relationship between a measure of the LGD systematic risk and a price of risk on average. Witzany et al. (2012, p. 8) assume that one cannot recover more than the EAD, in which case LGD has a floor value of zero. The total discounted future recovery is assumed to be positive and LGD therefore capped at one. Put differently: LGD can therefore not be lower than zero or higher than one. The realised LGD for worked out accounts can be calculated, but for non-worked out accounts recovery data will not yet be available. The observed realised recovery rate on defaulted accounts can be used to estimate expected LGD for non-default accounts (Witzany et al., 2012, p. 8).

The EWSA model is based on a survival analysis model and we will discuss in the next section the general concept of survival analysis (Section 2.2) and in Section 2.3 the Cox proportional hazard model used in the EWSA model.

2.2 Survival analysis

To define the survival analysis used in the EWSA model we will first give a brief definition of survival analysis and then explain the concept of the survival function and the hazard rate and how these relate to LGD.

Survival analysis is generally defined as the set of procedures to study data where the end result is the time until the manifestation of a certain event (such as death or, in this instance, failure to repay a loan). Within this analysis some observations are censored. Censoring of observations takes place where observations survive up to a point in time, but where further information is unavailable. Where defaulted receivables are examined the elementary amounts are observed with individuals which are in the process of collection, up to the point where they repay. According to Kalbfleisch and Prentice (2002, pp. 6–7), Collet (2003, p. 11) and Greene (2003, pp. 903–904) the most important concepts central to the survival analysis methodology are the survival function and the hazard rate (Witzany et al., 2012, p. 13).

Where observations typically remain in a specific state until such a time as when a change occurs, survival analysis is an appropriate methodology. Survival analysis is typically used to view a person's mortality. However, in this article, the current state is referred to as survival and the exit point is where failure occurs. Observations that have survived with certainty will be censored where no more information is obtainable. In an LGD scenario the EAD is assessed in terms of whether it has survived the default state or not. A repayment can be interpreted as some of the EAD not having survived the default rate.

The survival function is defined as the probability of an event occurring after a specified time t such as that the EAD will remain in the default state. Witzany et al. (2012, p. 13) define the survival function as

$$S(t) = 1 - F(t) = 1 - P(T < t),$$

where the random variable T denotes the time of the event and the cumulative distribution function is denoted as $F(t)$. $S(t)$ and $F(t)$ respectively give the expected loss rate at t and the expected recovery rate given that the process terminates at t. The corresponding probability density function is $f(t)$. The hazard rate, $h(t) = f(t)/S(t)$, is the instantaneous rate of exit at t, given that survival has been attained up to point t. In an LGD setting the hazard rate is the instantaneous rate of recovery at point t given that the EAD survived default up to point t. The probability that the EAD exits default in the time interval $(t, t + \Delta t]$, given that $EAD_{i,\mathrm{t}}$ is still in default at t, is $h(t)\Delta t$. The survival function, $S(t) = e^{-\mathrm{H(t)}}$, is expressed in terms of the cumulative hazard function $\mathrm{H}(t) = \int_0^t \lambda(t)ds$ (Witzany et al., 2012, pp. 13–14).

There are two options to define the hazard rate: the parametric and semi-parametric methods. In this paper, the semi-parametric method is adopted, which is described in the following section.

2.3 The Cox proportional hazard model

Witzany et al. (2012, pp. 13–14) define the semi parametric Cox proportional hazards model as

$$h(t, \mathbf{x}) = h_0(t) \exp(\mathbf{x}'\beta),$$

with the 0 emphasising that $h_0(t)$ is the *baseline* hazard. The baseline hazard is independent of the covariate values, $\mathbf{x}$. The matching survival function is

$$S(t, \mathbf{x}) = \exp\left(-\int_0^t h_0(s)\exp(\mathbf{x}'\beta)\, ds\right) = S_0(t)^{\exp(\mathbf{x}'\beta)},$$

where $S_0(t) = \exp(-\int_0^t h_0(s)ds)$. The partial likelihood is used to estimate the parameter vector β. The partial likelihood for a specific account i that exits at time t is defined as

$$L_i(\beta) = \frac{h(t, \mathbf{x}_i)}{\sum_{j \in A_i} h(t, \mathbf{x}_j)} = \frac{\exp(-\mathbf{x}_i'\beta)}{\sum_{j \in A_i} \exp(-\mathbf{x}_j'\beta)},$$

with $\mathbf{x}_i$ the set of covariates at the point of exiting default and A_i the set of objects in default at t. It is assumed that there is only one exit at time t. Given that there are K accounts, the equation

$$\ln(L) = \sum_{i=1}^{K} \ln(L_i)$$

is maximised by using the Newton–Raphson algorithm to obtain the estimate for β. When modelling LGD, multiple exits may occur and the partial likelihood is adapted to handle ties. An approximation of the partial likelihood is used to solve the parameter estimates in the case where ties occur. The baseline hazard function is assumed to be constant for each unit time interval and are estimated separately. The likelihood function

$$L_t = \prod_{i=1}^{n} [h_0(t)\exp(\mathbf{x}_i'\beta)]^{dN_i(t)} \exp\left(-h_0(t)\exp(\mathbf{x}_i'\beta)Y_i(t)\right)$$

is then maximised. Each indicator $Y_i(t)$ indicates that observation i has not exited default at $t-1$ and is incomplete. Each indicator $dN_i(t)$ indicates that observation i exited from default at $(t-1, t]$ by curing or writing off. The Breslow–Crowley form for the maximum likelihood estimator of the baseline hazard is then

$$\widehat{h}_0(t) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} \exp(\mathbf{x}_i'\beta)Y_i(t)}$$

(Witzany et al., 2012, pp. 13–14).

2.4 Survival analysis in LGD modelling

In order to tailor the survival analysis methodology, we need to make a few assumptions and need to construct the data. This section outlines the assumptions and data collection methodology.

All accounts with recovery information up until time T_w are deemed to have complete recovery information. The time when the recovery process ends for account i will be denoted by $t_{i,end}$. The recovery process is completed if $t_{i,end} < T_w$, for example, if an account closes before T_w. Alternatively, the recovery process is incomplete if $T_w \leq t_{i,end}$. The constructed dataset contains not only a record for each recovered amount, but also a record for each amount not recovered. This can be constructed in the following way. First create a record for each discounted recovered amount (cash flow) that is positive, $DCF_{i,t}$. A positive recovery indicates that a portion of the EAD exited the default state. Represent this recovery by creating an observation containing a frequency weight equal to the $DCF_{i,t}$ amount. To accommodate censoring we need to create a censoring variable. Let the censoring variable be equal to zero to indicate that we are dealing with an exit event. The amount in the frequency weight is recovered and is exiting default. This record is created at the time that the recovery occurs, t. The unrecovered amount is calculated as

$$d_i = EAD_{i,0} = \sum_{t=1}^{T_w} DCF_{i,t}.$$

Create a record with a frequency weight equal to the unrecovered amount d_i. Let the censoring variable be equal to 1 for an unrecovered amount, meaning that this amount did not exit default and is unrecovered. The time of this entry will differ depending on whether the unrecovered amounts are complete or incomplete. Create the record at time $t_{i,end}$ in the case where the recovery process is incomplete at $t_{i,end}$. If the recovery point is complete at $t_{i,end}$ the records need to be created at time T_w at the end of the recovery process (Witzany et al., 2012, pp. 16–17).

An example of data constructed for the survival analysis methodology is illustrated in Table 1. A recovery of $DCF_{i,1} = 100$ is made at $t = 1$ on account 14. A record with a frequency weight equal to 100 and a censoring variable of 0 is created at time $t = 1$. A similar record is created for every other recovery made on account 14. The record created for the last recovery on account 14 has a frequency weight of $DCF_{i,T_w} = 148$ and a censoring variable of 0 that is created at point $t = T_w$. Account 14 has complete recovery information since the unrecovered amount of $d_i = 123$ occurs at $t_{end} = T_w$. A record with a frequency weight equal to 123 and a censoring variable of 1 is created. Account 15 has a recovered amount of $DCF_{i,1} = 300$ at $t = 1$. A record with a frequency weight equal to 300 is created at point $t = 1$ and the censoring variable has a value of 0. A similar record is created for every other recovery made on account 15. The last recovery on account 15 occurs at $t = 8$ and the value of this recovery is $DCF_{i,1} = 169$. The record for the over recovery that is created at $t = 8$ will have a frequency weight of 169 and a censoring variable of 1. There is no further information on this account and the account is deemed to be incomplete at this point. The record for the unrecovered amount is therefore created at point $t = 8$ with a frequency weight of $d_i = 256$ and a censoring variable of 1.

A survival curve $S(t,i) = 1 - P(T < t)$ is defined as the (unrecovered) proportion of $EAD_{i,0}$ that remains in default up to a specific recovery time t, where $t \in \{1, \ldots, T_w\}$ for account i. Thus

$$S(t,i) = \frac{EAD_{i,0} - \sum_{s=1}^{t} DCF_{i,s}}{EAD_{i,0}}.$$

The Kaplan–Meier estimate $\widehat{S}(t,i)$ is the empirical value of the survival curve calculated from the data and is equal to

$$\widehat{S}(t,i) = \frac{\widehat{EAD}_{i,0} - \sum_{s=1}^{t} \widehat{DCF}_{i,s}}{\widehat{EAD}_{i,0}},$$

where $\widehat{EAD}_{i,0}$ is the exposure at the default point for account i and $\widehat{DCF}_{i,s}$ is the value of the cash flow for account i at point s. The survival curve for the population is then calculated as

$$\widehat{S}_0(t) = \frac{\widehat{EAD}_0 - \sum_{s=1}^{t} \widehat{DCF}_s}{\widehat{EAD}_0},$$

Table 1. Example of a dataset for survival analysis.

Account i	t	Frequency weight	Censoring
14	1	100	0
14	2	89	0
14	3	0	0
⋮	⋮	⋮	⋮
14	8	158	0
⋮	⋮	⋮	⋮
14	T_w	148	0
14	T_w	d_i = 123	1
15	1	300	0
15	2	400	0
15	3	155	0
⋮	⋮	⋮	⋮
15	8	169	0
15	8	d_i = 256	1
⋮	⋮	⋮	⋮

where $\widehat{EAD}_0 = \sum_i \widehat{EAD}_{i,0}$ and $\widehat{DCF}_s = \sum_i \widehat{DCF}_{i,s}$.

The general form of the Cox model can be written as

$$S(t, x_i) = S_0(t)^{\exp(\mathbf{x}_i'\beta)}.$$

The weighted survival curve at time t in default contains a component $S_0(t)$ that is known as the baseline survival curve. The baseline survival curve is the Kaplan–Meier estimate of the portfolio where the dummy variables are equal to the base group. If an account falls outside of the base of the dummy variable, the baseline $S_0(t)$ is shifted by the exponent of $\exp(\mathbf{x}_i'\beta)$. The loss given default for account i at point t in default is calculated as

$$LGD_{i,t} = \frac{S(T_w, \mathbf{x}_i)}{S(t, \mathbf{x}_i)}.$$

3. Default weighted survival analysis (DWSA)

The main contribution of this article is the following three enhancements that were made to the EAD weighted methodology by Witzany et al. (2012): over-recoveries, default weighted and negative cash flows.

3.1 Over-recoveries

Witzany et al. (2012) did not cater for over-recoveries, which will occur in practice when the expected amount recovered is more than the EAD, i.e. $\sum_{t=1}^{T_w} DCF_{i,t} > EAD_{i,0}$. In this article, a technique to cater for over-recoveries will be included. In the following example we will explain how over-recoveries are accommodated for in the algorithm. In Table 2 we give an example of three accounts with recovered discounted cash flows and EAD.

Table 2. Loss given default example.

		Discounted cash flows Months since default		
	EAD	**1**	**2**	**3**
Account A	100	20	0	60
Account B	250	150	320	0
Account C	320	180	10	18
Total	670	350	330	78

The LGD for the portfolio is calculated as

$$LGD_0 = \frac{\widehat{EAD}_0 - \sum_{s=1}^{t} \widehat{DCF}_s}{\widehat{EAD}_0} = \frac{670 - (350 + 330 + 78)}{670} = -13.134\%,$$

where

$$\widehat{EAD}_0 = \sum_i \widehat{EAD}_{i,0} \qquad \text{and} \qquad \widehat{DCF}_s = \sum_i \widehat{DCF}_{i,s}.$$

The Kaplan–Meier estimate of the survival curve is

$$S(t) = \frac{\widehat{EAD}_0 - \sum_{s=1}^{t} \widehat{DCF}_s}{\widehat{EAD}_0}$$

and is calculated from the data as:

$$\widehat{S}(1) = \frac{\widehat{EAD}_0 - \sum_{s=1}^{1} \widehat{DCF}_s}{\widehat{EAD}_0} = \frac{670 - 0}{670} = 1,$$

$$\widehat{S}(2) = \frac{\widehat{EAD}_0 - \sum_{s=1}^{2} \widehat{DCF}_s}{\widehat{EAD}_{i,0}} = \frac{670 - 350}{670} = 47.76\%,$$

$$\widehat{S}(3) = \frac{\widehat{EAD}_0 - \sum_{s=1}^{3} \widehat{DCF}_s}{\widehat{EAD}_0} = \frac{670 - 350 - 330}{670} = -1.49\%,$$

$$\widehat{S}(4) = \frac{\widehat{EAD}_0 - \sum_{s=1}^{4} \widehat{DCF}_s}{\widehat{EAD}_0} = \frac{670 - 350 - 330 - 78}{670} = -13.13\%.$$

The empirical survival curve values are negative for months on book equal to three and four. Traditional survival analysis does not allow for negative empirical values on a survival curve and the proportional hazards procedure in SAS software will only cater for survival curves with positive empirical values.

In order to accommodate the over-recoveries the unrecovered amount ($EAD_{i,0} - \sum_{s=1}^{t} DCF_{i,s}$) at each recovery time is adjusted upwards in such a way that the resulting values of the empirical survival curve will be positive. Typical survival analysis software (e.g., the proportional hazards procedure in SAS) will now be able to fit the survival curve with positive values. The effect of this adjustment will be reversed and the original survival curve, which contains negative empirical values, obtained.

The adjusted value is simply the maximum over recovered amount of all the accounts. Thus we define the maximum over recovered amount OR as:

$$OR = \max_i \left(\widehat{EAD}_{i,0} - \sum_{s=1}^{T_w} \widehat{DCF}_{i,s} \right).$$

Therefore $S(t,i)$ is updated as

$$S^*(t,i) = \frac{EAD_{i,0} - \sum_{s=1}^{t} DCF_{i,s} + OR}{EAD_{i,0}}.$$

$S^*(t,i)$ is calculated by making use of the proportional hazards procedure in SAS software. The inflated survival curve for the example is calculated and given in Table 3. The values that are required to adjust the inflated survival curve $S^*(t,i)$ back to its original values are the month on month inflated recovery rate $MR^*(t)$ as defined by

$$MR^*(t,i) = 1 - \frac{S^*(t,i)}{S^*(t-1,i)}.$$

We define the inflated exposure ratio $R^*(t,i)$ as

$$R^*(t,i) = \frac{EAD_{i,0} - \sum_{s=1}^{t} DCF_{i,s} + OR}{EAD_{i,0} - \sum_{s=1}^{t} DCF_{i,s}}.$$

To obtain the month on month recovery rate $MR(t,i)$ the month on month inflated recovery rate is multiplied by the inflated exposure ratio,

$$MR(t,i) = MR^*(t,i) \times R^*(t,i).$$

The values for the survival curve are then updated as

$$S(t,i) = S(t-1,i) - S(t-1,i) \times MR(t,i).$$

For the empirical portfolio survival curve $\widehat{S}_0(t)$ the subscript i is dropped and we take $\widehat{EAD}_0 = \sum_i \widehat{EAD}_{i,0}$ and $\widehat{DCF}_s = \sum_i \widehat{DCF}_{i,s}$, then $\widehat{S}_0(t) = \widehat{S}_0(t-1) - \widehat{S}_0(t-1) \times \widehat{MR}(t)$.

The values for $\widehat{MR}(t)$, $\widehat{MR}^*(t)$, $\widehat{R}^*(t)$, $\widehat{S}_0^*(t)$ and $\widehat{S}_0(t)$ for the example in Table 2 are given in Table 3. The empirical loss given default value is calculated as

$$\widehat{LGD}_0 = \frac{\widehat{S}_0(T_w)}{\widehat{S}_0(0)} = \frac{-13.13\%}{100\%} = -13.13\%.$$

This value is equal to the empirical loss given default that is calculated from Table 3.

3.2 Default-weighting

The methodology by Witzany et al. (2012) results in a $EAD_{i,0}$ weighted $LGD_{i,t}$ estimate. The Basel accord states that LGD cannot be less than the long run default weighted average loss rate (BCBS, 2006, p. 103). An approach to estimate the default weighted $LGD_{i,\mathrm{t}}$ estimates will be developed and described.

Table 3. Over recovery adjustments.

	Months since default			
	0	**1**	**2**	**3**
Unrecovered amount	670	320	-10	-88
Unrecovered amount + OR	890	540	210	132
$\widehat{S}_0^*$(t)	100.00%	60.67%	23.60%	14.83%
$\widehat{MR}^*(t)$		39.33%	61.11%	37.14%
$\widehat{R}_t^*$	132.84%	168.75%	-2100.00%	-150.00%
$\widehat{MR}(t)$		52.24%	103.13%	-780.00%
$\widehat{S}_0(t)$	100.00%	47.76%	-1.49%	-13.13%

The data are constructed in a specific way when applying the DWSA methodology. The dataset contains a record for every recovery made and for every recovery that is not made. The recovery made on account i at point t in default is discounted to the default point. In the case of the ESWA methodology a record with a frequency weight equal to the discounted recovery $DCF_{i,t}$ is created. For the DWSA approach, the discounted recovery $DCF_{i,t}$ is divided by $EAD_{i,0}$ to obtain the default weighted discounted recovery $DCF_{i,t}/EAD_{i,0}$. The value of the frequency vector for the DWSA approach is set equal to the default weighted discounted recovery. The censoring variable for this record will be equal to zero to indicate that we are dealing with an exit event. This record will be created at the time that the cash flow takes place. The unrecovered amount for the EWSA approach equals

$$d_i = EAD_{i,0} - \sum_{t=1}^{T_w} DCF_{i,t} \, ,$$

with T_w the last possible point for a recovery to take place. The unrecovered amount is divided by the exposure at the default point to obtain the default weighted unrecovered amount $d_i/EAD_{i,0}$. A record with a frequency weight equal to the default weighted unrecovered amount will be added to the dataset for the DWSA approach. The censoring variable will be equal to one to indicate an unrecovered amount. The timing of this record will differ depending on whether the unrecovered amount is used in the calculation of the default weighted unrecovered amount is complete or incomplete. A record is deemed complete if the record contains recovery information up until T_w and the record for a complete recovery will be created at T_w. The record can be complete or incomplete if no further recovery for a record is available from point t_{end} onwards depending on the reason for the missing information. A closed account will have no further information from point t_{end} onwards, but is deemed complete and the record created at T_w. The record for an incomplete account is created at point t_{end} (Witzany et al., 2012, pp. 16–17).

The actual value of the survival curve $\widehat{S}_t$ calculated from the data is equal to

$$\widehat{S}(t,i) = \frac{\widehat{EAD}_{i,0} - \sum_{s=1}^{t} \widehat{DCF}_{i,s}}{\widehat{EAD}_{i,0}} .$$

The Cox proportional hazards model is

$$S\left(t, \mathbf{x}_i\right) = S_0(t)^{\exp(\mathbf{x}_i' \beta)},$$

with $S_0(t)$ the baseline survival curve. The baseline survival curve is the Kaplan–Meier estimate for the base population. The base population is the population where the dummy variables are equal to the base group. This baseline survival curve $S_0(t)$ is adjusted by $\exp(\mathbf{x}_i'\beta)$ when the covariates fall outside of the baseline group. The loss given default for account i at point t is

$$LGD_{i,t} = \frac{S(T_w, \mathbf{x}_i)}{S(t, \mathbf{x}_i)}.$$

The above-mentioned default weighted $LGD_{i,t}$ is averaged over an extended period to produce the long run default weighted average loss rate. The Basel accord stipulates that the LGD that is used to calculate regulatory capital should not be less than this long run default weighted average loss rate (BCBS, 2006, p. 103).

3.3 Negative cash flows

A description of the cash flow calculation is given in Section 2.1. In practice negative cash flow will occur in the form of recovery process cost such as legal and administrative costs. The negative recoveries in the EWSA approach are set to zero. The LGD will be underestimated if these recovery process costs are not included in estimating LGD. A technique to include negative cash flows into LGD modelling under the DWSA approach will be developed and described.

The EWSA approach that is used by Witzany et al. (2012) sets all negative cash flows to zero, $DCF_{i,t} = 0$ when $DCF_{i,t} < 0$. For each defaulted account, negative cash flows may be contained within the observed cash flow stream. As survival analysis cannot cater for negative cash flows, the adjustment methodology for the DWSA approach is explained.

Two separate datasets will be constructed. In the first dataset, all negative cash flows will be set to zero. A record for every recovery is created with the frequency variable equal to the default weighted discounted recovery $DCF_{i,t}/EAD_{i,0}$, and the negative recoveries are set to zero, $DCF_{i,t} = 0$ if $DCF_{i,t} < 0$. The censoring variable will be equal to zero to indicate that a recovery is made. This record will be created at time t where the cash flow takes place. A record is created for every unrecovered amount. The frequency variable for the unrecovered amount will equal

$$\frac{d_i}{EAD_{i,0}} = \frac{EAD_{i,0} - \sum_{s=1}^{T_w} DCF_{i,s}}{EAD_{i,0}},$$

and the negative cash flows will be set to zero, $DCF_{i,t} = 0$ if $DCF_{i,t} < 0$. The censoring variable will be set to one to indicate an unrecovered amount. This record will be created at T_w for a complete record and created at t_{end} for an incomplete account. Create the positive survival curve $S^p(t)$ from the dataset where all negative cash flows are set to zero. The actual value of the survival curve $\widehat{S^p}(t)$ calculated from the data is equal to

$$\widehat{S^p}(t,i) = \frac{\widehat{EAD}_{i,0} - \sum_{s=1}^{t} \widehat{DCF}_{i,s}}{\widehat{EAD}_{i,0}},$$

where $DCF_{i,t} = 0$ if $DCF_{i,t} < 0$. The Cox proportional hazards model is

$$S^p(t, \mathbf{x}_i) = S_0^p(t)^{\exp(\mathbf{x}_i'\beta)},$$

with $S_0^p(t)$ the baseline survival curve. The second dataset is constructed by setting all the positive cash flows to zero and changing the signs of the negative cash flows by multiplying them with minus one. A record for every recovery is created with the frequency variable equal to the default weighted discounted recovery, $DCF_{i,t}/EAD_{i,0}$ where the positive recoveries are set to zero, $DCF_{i,t} = 0$ if $DCF_{i,t} \geq 0$ and the negative recoveries are made positive $DCF_{i,t} = -1 \times DCF_{i,t}$ if $DCF_{i,t} < 0$. The censoring variable will be equal to one and the time value of the record will be the time when the cash flow takes place. The frequency weight for the unrecovered amount is equal to

$$\widehat{S}^n(t,i) = \frac{\widehat{EAD}_{i,0} - \sum_{s=1}^{t} \widehat{DCF}_{i,s}}{\widehat{EAD}_{i,0}},$$

where $DCF_{i,t} = 0$ if $DCF_{i,t} \geq 0$ and $DCF_{i,t} = -1 \times DCF_{i,t}$ if $DCF_{i,t} < 0$. The Cox proportional hazards model is

$$S^n(t, \mathbf{x}_i) = S_0^n(t)^{\exp(\mathbf{x}_i'\beta)},$$

with $S_0^n(t)$ the baseline survival curve. Combining these two survival curves produces

$$S(t, \mathbf{x}_i) = S^p(t, \mathbf{x}_i) + 1 - S^n(t, \mathbf{x}_i),$$

and the loss given default can be calculated as

$$LGD_{i,t} = \frac{S(T_w, \mathbf{x}_i)}{S(t, \mathbf{x}_i)}.$$

Three enhancements were made to the EWSA methodology. By making use of default weighted survival analysis, the methodology was brought into closer alignment with the Basel requirements. By catering for negative cash flows and over-recoveries the modelling technique was brought into closer alignment with practice. The resulting DWSA methodology will be applied to the data described in the next section and is compared to alternative modelling techniques.

4. Data

The three datasets utilised in this paper are obtained from one of the big South African retail banks. These datasets are described in Section 4.1. In addition, we simulate five datasets which are described in Section 4.2. In the remainder of this article, various approaches to model LGD will be compared.

4.1 Retail banks datasets

A retail bank's credit card, revolving loan and cheque account datasets are used to compare various LGD modelling techniques. The three product sets chosen are derived from unsecured retail credit loan products from a large South African bank which are available for this study. The data has a significant history available and gives a good representation of a typical unsecured product within the South African context. The EAD, cash flows, discount rate, month of default and account number are stored monthly for each of these products. The cash flows are discounted to the default point and the loss given default calculated. The loss given default for the retail datasets is displayed in Figure 1. The cash flow values $CF_{i,t}$ include both positive and negative values in the actual data and the discounted cash flow is $DCF_{i,t} = CF_{i,t}/(1+r)^t$. The loss given default is calculated as

$$LGD_{i,0} = \frac{EAD_{i,0} - \sum_{t=1}^{T_w} DCF_{i,t}}{EAD_{i,0}},$$

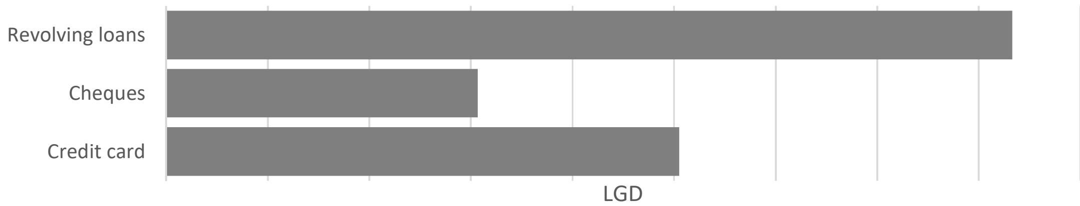


Figure 1. Retail banks datasets loss given default.

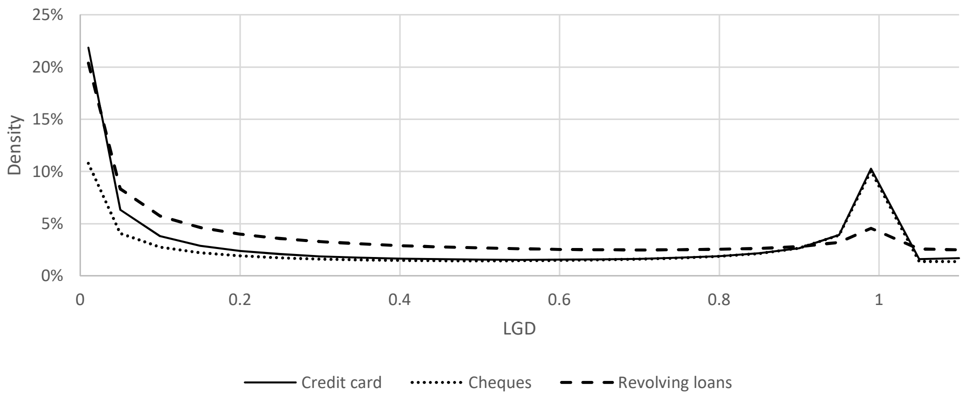


Figure 2. Retail banks datasets LGD distribution.

for account i at time t. The loss given default values that are displayed in Figure 1 consist of both negative and positive cash flow values. The LGD axis in all figures is left out due to confidentiality.

The percentage of negative cash flows for each of the respective datasets are 1.89%, 2.17% and 1.72%. The distribution of the loss given default for the retail banks datasets is displayed in Figure 2. Figure 2 shows that over-recoveries are present on these datasets. Over-recoveries occur where the loss given default value is greater than one.

Variables used in this study are selected from the following main data categories: behavioural, application, customer, bureau, demographic and macroeconomic. The reference period for all the development datasets ranges from December 2007 to November 2009. The period used was determined by using the representative economic downturn conditions as required by Basel (BCBS, 2006, par. 468). The highest twenty-four month average losses occurred during the stated period. The three development datasets respectively have 90 691, 22 300 and 55 983 accounts.

The actual values for the hazard rate, distribution and survival curve for the positive, negative and combined cash flows are displayed for the credit card, revolving loan and cheque account datasets. The top left graph in Figure 3 displays the actual values of the hazard rate and the distribution for the entire credit card portfolio. The top right graph displays the actual values for the hazard rate and the probability where only positive cash flow values are included. The bottom left graph contains the actual values for the hazard rate and the distribution where only negative cash flows are included. The bottom right graph contains the survival curves for the total population, positive cash flows and negative cash flows. The same layout is repeated in Figure 4 and Figure 5 for the empirical values of the cheque data and revolving loan data, respectively.

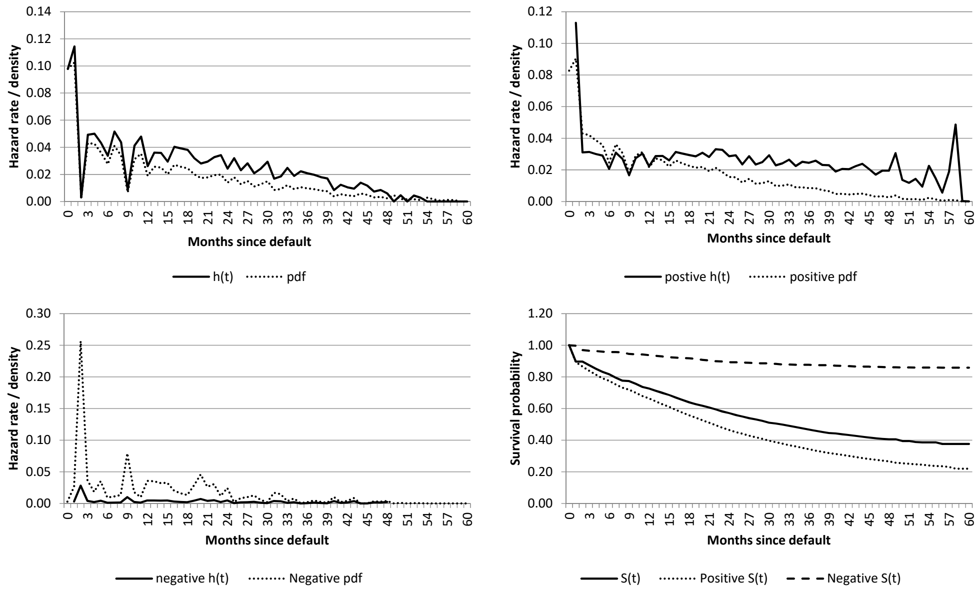


Figure 3. Credit card actual hazard rate, distribution and survival curve.

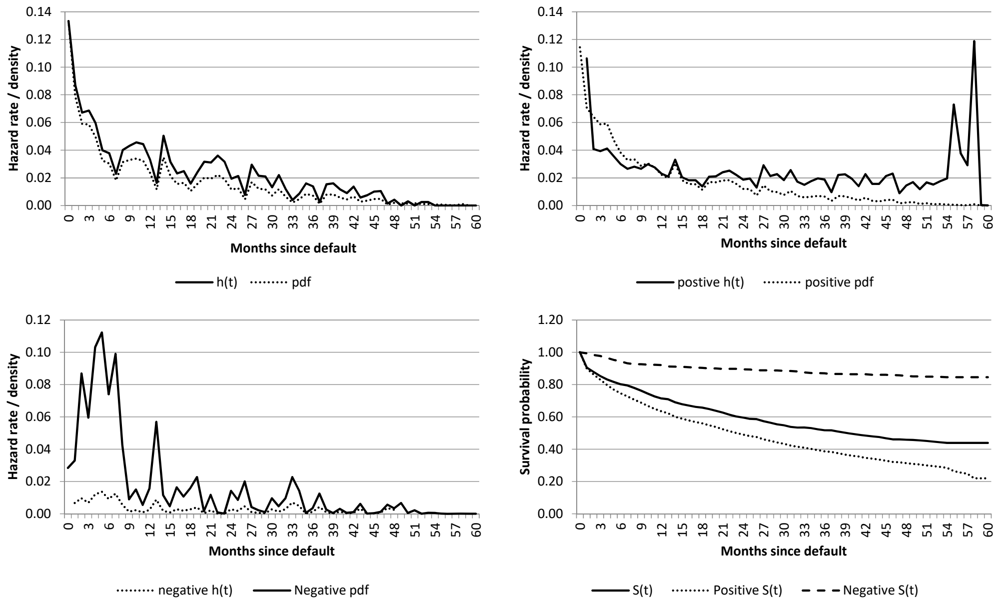


Figure 4. Cheque actual hazard rate, distribution and survival curve.

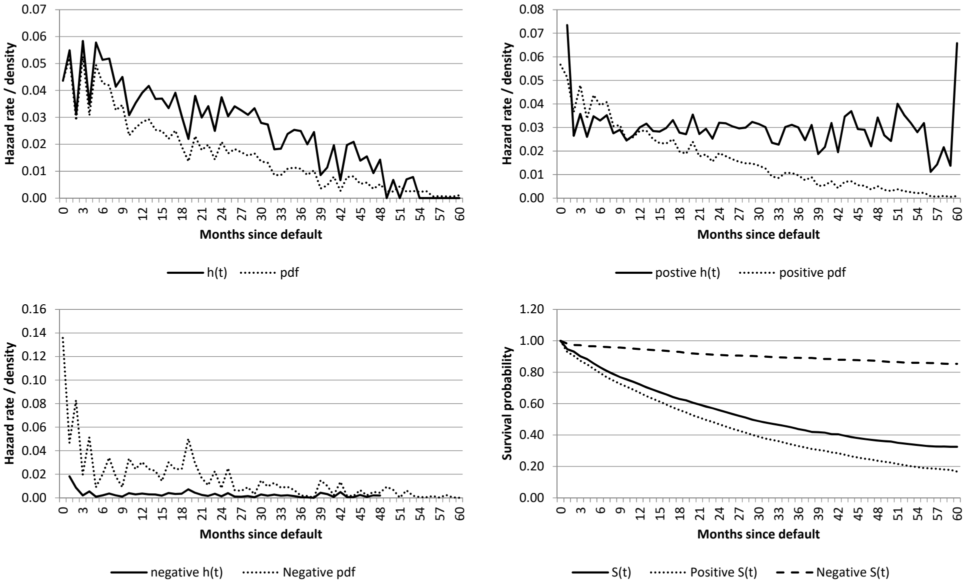


Figure 5. Revolving loan actual hazard rate, distribution and survival curve.

4.2 Simulated datasets

To construct the simulated datasets we use the beta distribution to simulate recovery rates and the gamma distribution to simulate EAD. The beta distribution is widely used to simulate recovery rates (Chen and Wang, 2013, p. 1). The EAD is simulated from a gamma distribution (Jimenez and Mencia, 2009, p. 8). These values are simulated for every account.

A workout process of 60 months is assumed and a random uniform number between 0 and 60 is generated to represent the point at which an account will exit default. A record is created for every point at which an account is in default. The first record is where the account is zero months since default and the months since default variable is populated until the account exits default. A random number is used to assign the total recoveries for an account to the various months that account is in default. This random number can take a positive or a negative value and the assigned monthly recovery values can therefore be positive or negative. For every account, the sum of the monthly recoveries is equal to the total recovery simulated for that account. The percentage of negative cash flows simulated for each of the respective simulated datasets are 1.74%, 2.17%, 1.72%, 1.79% and 2.02%. Some individuals recover by paying the same amount every month, others start off by paying bigger amounts which then become less over time. There are many differences in how recoveries are structured and therefore a uniform random variable was used to indicate how many months there are to recovery and then randomly determine how many recoveries would be collected. Since all the recoveries are aggregated at the end of the period, the effect of the type of time-recovery is assumed to be negligible and was omitted for this paper. The investigation of this effect is for future research.

The process to determine the beta distributions parameter estimates, used in the recovery rates simulation, is described next. Beta distributions are fitted to various retail bank portfolios and the parameter estimates are illustrated in Figure 6. Parameter estimates in the same range are used in the simulation study.

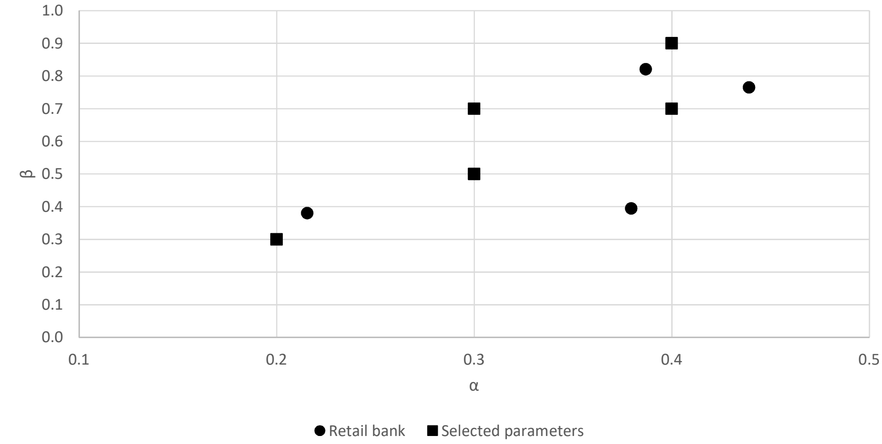


Figure 6. Beta distribution parameter estimates.

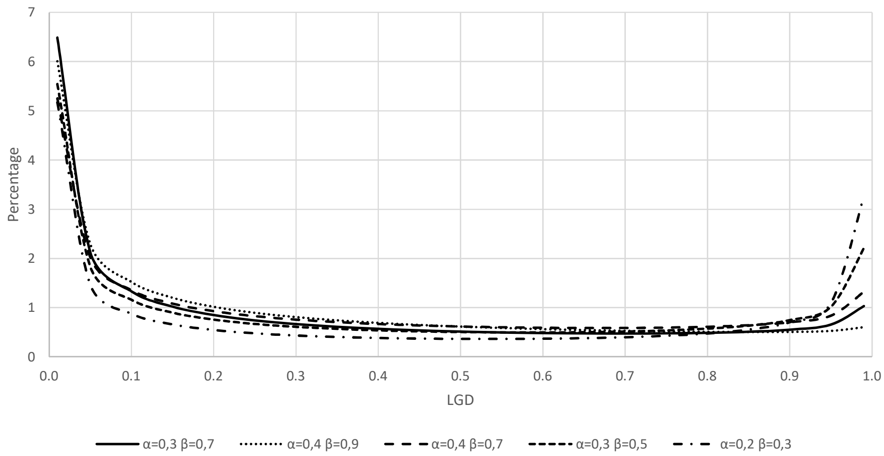


Figure 7. Beta distribution pdf.

The probability density function (pdf) of the beta distributions,

$$f\,(x;\alpha;\beta) = \frac{\Gamma\,(\alpha+\beta)}{\Gamma\,(\alpha)\,\Gamma\,(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1,\, \alpha,\beta > 0,$$

for each pair of parameter estimates, α and β, used in the simulation is displayed graphically in Figure 7.

The parameters in Figure 6 give rise to pdfs as displayed in Figure 7. In Figure 7 one can deduce that an LGD of close to zero (lower LGD value) occurs frequently and that an LGD of close to one (higher LGD value) occurs frequently in some instances and infrequently in other instances. LGD values between the lower and higher LGD values have a constant frequency. The form of these simulated pdfs is typical of retail banks in South Africa. The distribution of the overall loss given default is displayed in Figure 7. The loss given default has a similar shape and outcome than that of the graph in Witzany et al. (2012, p. 20). The reality therefore matches expectations around full recovery. Over-recoveries are artificially added to the dataset used to produce Figure 7 resulting in

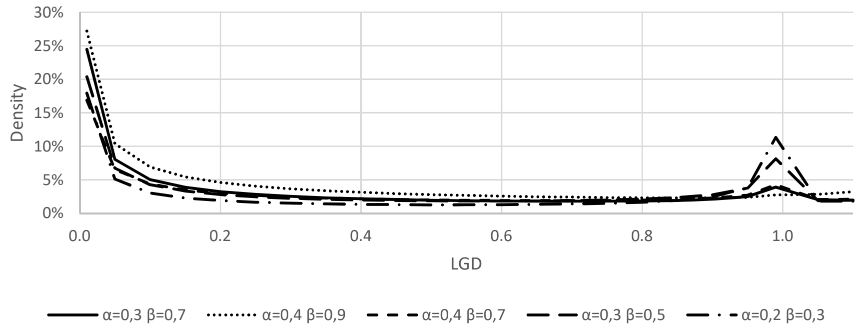


Figure 8. Beta distribution pdf with artificially added over-recoveries.

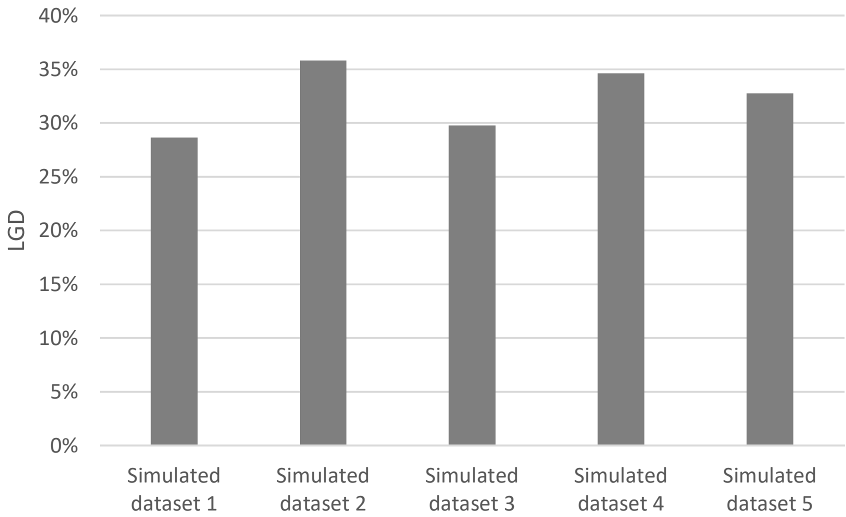


Figure 9. Simulated datasets loss given default.

Figure 8. The actual simulated LGD values are displayed in Figure 9 and correspond to the datasets displayed in Figure 8.

The parameter estimates of a gamma distribution are used to estimate the EAD in the simulation. Gamma distributions are fitted to the EAD for retail bank portfolios. The values of the gamma distribution for these portfolios are graphically displayed in Figure 10. Again parameter estimates in the same range are used in the simulation study. The probability density function of the gamma distributions is

$$f(x; k; \theta) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad x > 0,\ k > 0,\ \theta > 0,$$

and is graphically displayed in Figure 11 for the parameter estimate used to simulate the EAD.

The beta parameters and gamma parameters that are used in the simulations are given in Table 4.

The hazard rate distribution and survival curves for each of the simulated datasets are displayed for the positive cash flow values, negative cash flow values and the total population.

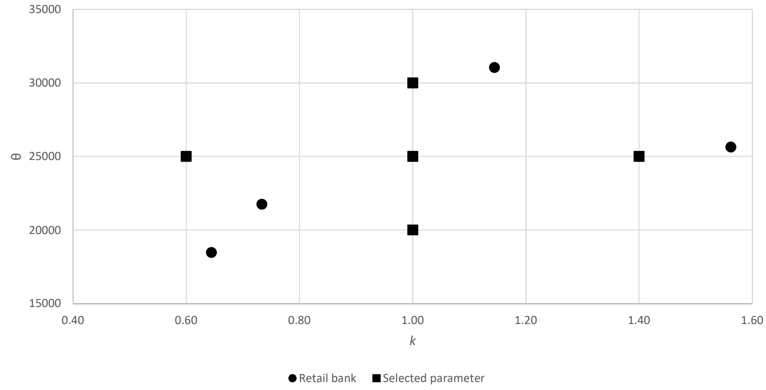


Figure 10. Gamma distribution parameter estimates.

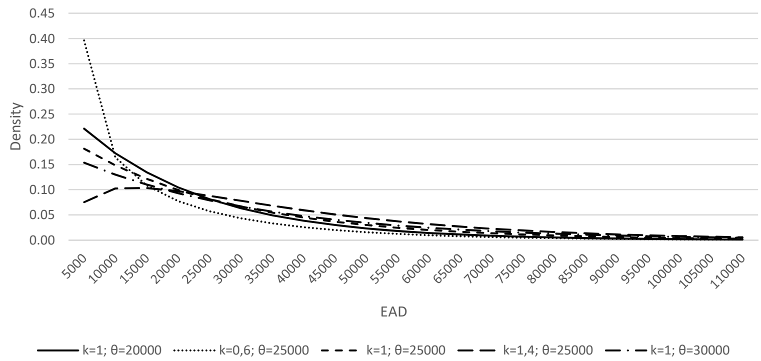


Figure 11. Gamma distribution pdf.

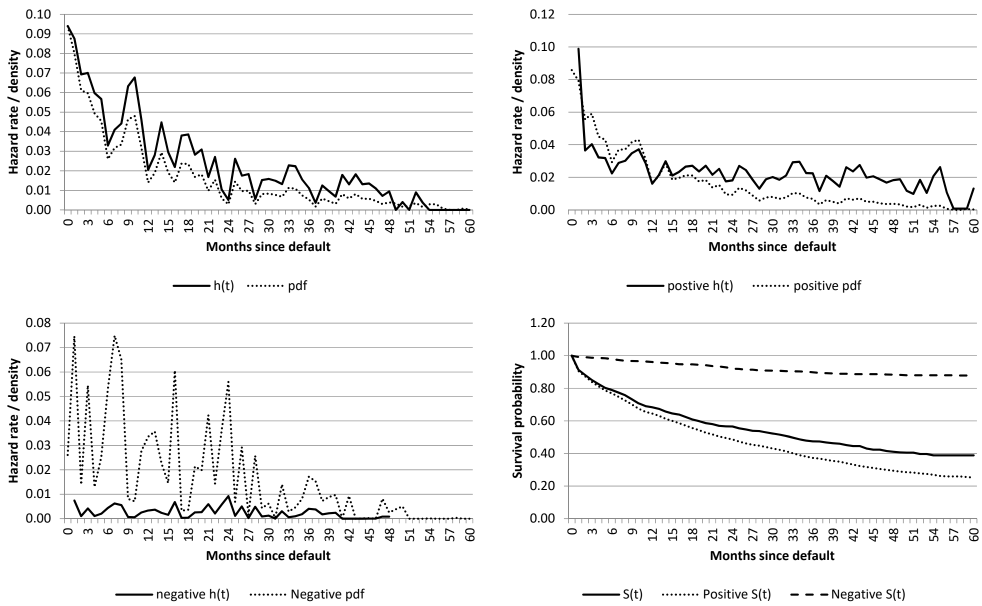


Figure 12. Simulated dataset 1 actual hazard rate, distribution and survival curve.

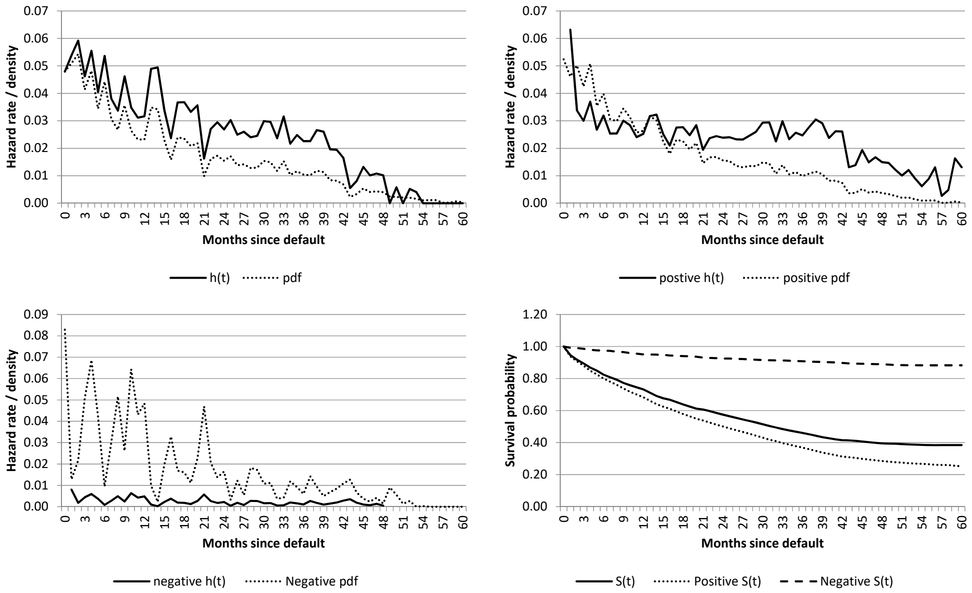


Figure 13. Simulated dataset 2 actual hazard rate, distribution and survival curve.

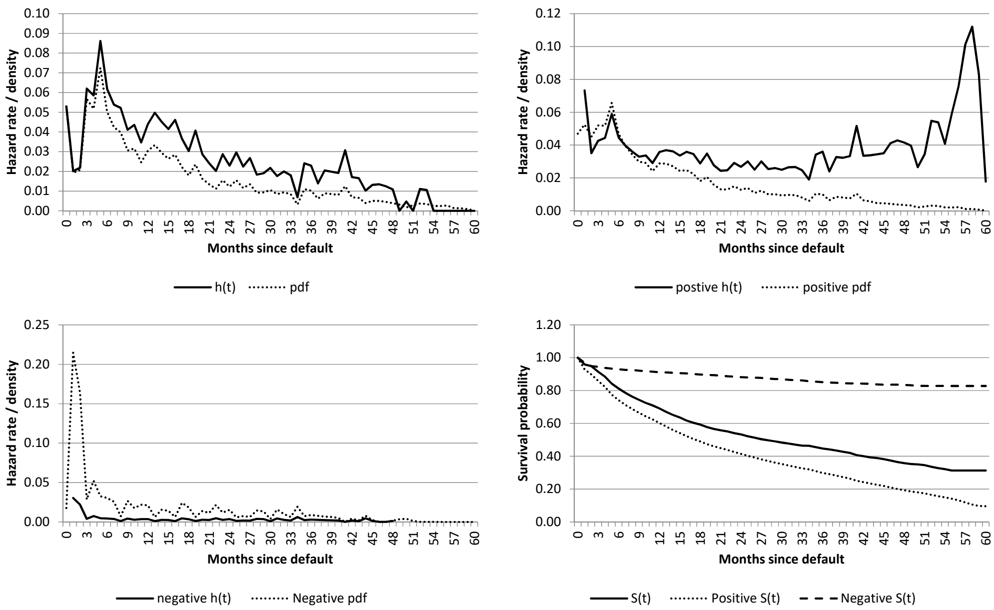


Figure 14. Simulated dataset 3 actual hazard rate, distribution and survival curve.

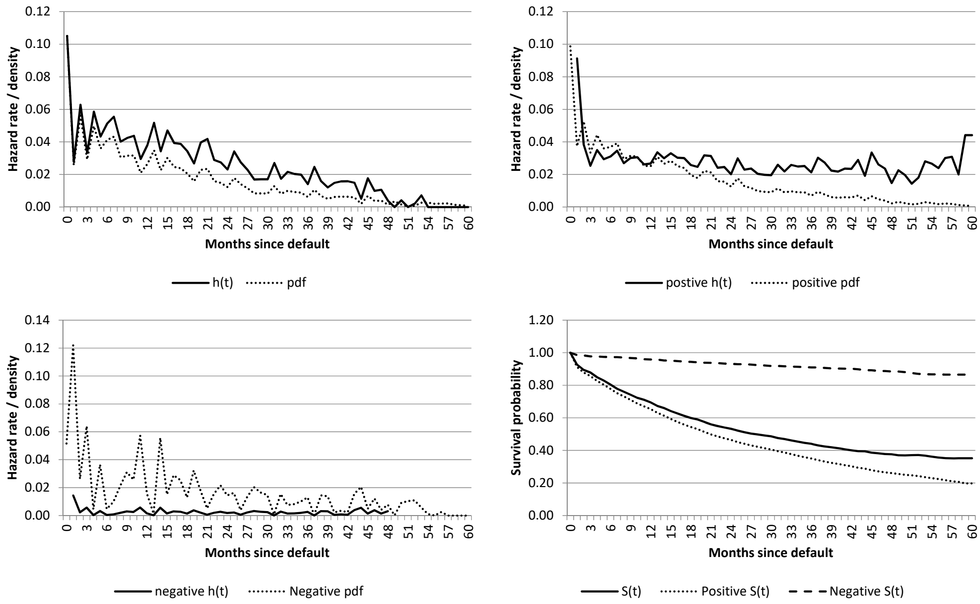


Figure 15. Simulated dataset 4 actual hazard rate, distribution and survival curve.

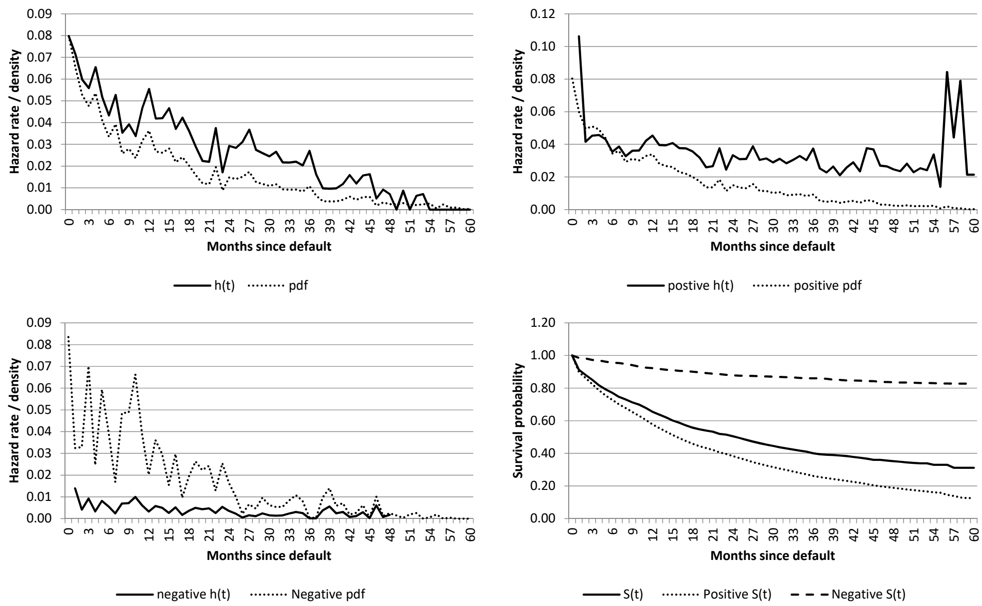


Figure 16. Simulated dataset 5 actual hazard rate, distribution and survival curve.

Table 4. Beta and gamma parameter estimates used in simulation.

	Beta parameters used	**Gamma parameters used**
Simulated data set 1	$\alpha = 0.2, \beta = 0.3$	$k = 1.0, \theta = 20\,000$
Simulated data set 2	$\alpha = 0.3, \beta = 0.5$	$k = 1.0, \theta = 25\,000$
Simulated data set 3	$\alpha = 0.3, \beta = 0.7$	$k = 1.4, \theta = 25\,000$
Simulated data set 4	$\alpha = 0.4, \beta = 0.7$	$k = 1.0, \theta = 30\,000$
Simulated data set 5	$\alpha = 0.4, \beta = 0.9$	$k = 0.6, \theta = 25\,000$

4.3 Model fit

The following approaches to model LGD directly are used in this paper: beta regression, ordinary least squares, fractional response regression, inverse beta, run-off triangle and Box–Cox model. These are compared to the survival analysis approach namely the EAD weighted survival analysis approach (EWSA) and the enhancements made to the default weighted survival analysis (DWSA) technique by Witzany et al. (2012). The Cox proportional hazards model is used to fit the DWSA and EWSA model. Descriptions of the beta regression, ordinary least squares, fractional response regression, inverse beta, run-off triangle and Box–Cox model are given in Appendix A. Data are simulated and the models are applied to both the retail and simulated data. The mean squared error, bias and variance are calculated.

4.3.1 Mean squared error, bias and variance

The mean squared error is equal to the squared bias plus the variance:

$$MSE = \mathrm{Var}\left(\widehat{LGD_{i,0}}\right) + \mathrm{Bias}\left(\widehat{LGD_{i,0}}, LGD_{i,0}\right)^2,$$

with $\widehat{LGD_{i,0}}$ the actual value of the LGD, calculated as

$$\widehat{LGD_{i,0}} = \frac{\widehat{EAD}_{i,0} - \sum_{t=1}^{T_w} \widehat{DCF_{i,t}}}{\widehat{EAD}_{i,0}}.$$

The expected value of the LGD is obtained from the model. As an example, the expected value for the DWSA LGD is expressed as

$$LGD_{i,0} = \frac{S(T_w, \mathbf{x}_i)}{S(0)},$$

where $S(0) = 1$.

5. Results

Previous studies made use of the EAD weighted survival analysis method (EWSA) and the main aim of this study is to improve on it by default weighting the LGD estimates, including negative cash flows into LGD modelling and catering for over-recoveries. The secondary aim of this study is to compare eight techniques to model LGD.

5.1 Retail bank datasets

In each of the retail datasets used, i.e. credit card, revolving loan and cheque, the account level expected LGD and actual LGD, defined in the Section 4.3 above, are used to calculate the account level MSE. The portfolio average MSE values are displayed in Figure 17.

When considering all three data sets, the results of the default weighted survival analysis (DWSA) model, displayed on the far left of Figure 17, yield the best result. Judging by the MSE the survival analysis displays the best fit. Not only does this model result in the lowest MSE, but it also displays the lowest bias and lowest variance. Despite the DWSA model outperforming all other models, the beta regression also performed well. The MSE for the beta regression is on average 2.08% higher than that of the DWSA method, when compared over the three retail datasets. The default weighted survival analysis (DWSA) method yielded favourable results in that the MSE is significantly lower than that estimated by the EWSA model. The improvements made therefore aid in estimating the LGD more accurately. It is interesting to note that the survival analysis method yields the lowest MSE on the cheques data, whereas all the other models yield the lowest MSE on the revolving loan data. Run-off triangles are traditionally used in practice; however, they underperform all other methods used in this comparison, except for the Box–Cox transformation, which performs the worst. Note that the squared bias is included in Figure 17, but due to low squared bias values, it is not always visible. Figure 18 displays the bias in more detail.

The bias is calculated by taking the difference between the actual LGD value and the expected LGD value. The difference between the actual LGD and the expected LGD is smallest when the DWSA model is used. The DWSA model yields the lowest bias on all three retail products. The average bias on the DWSA model is -1.11%. The bias of the beta regression model is on average 2.7%, putting it in second place when comparing the bias. The fractional response regression averages 3.8% (with a range of 3.4% to 4.53%). Other models deliver much higher bias values. The bias for the EWSA model is much higher than that of the DWSA model. The main cause for this difference is that the EWSA model sets the negative cash flows to zero and that this model does not cater for over-recoveries.

5.2 Simulated datasets

One hundred thousand (100 000) accounts are simulated for each of the five simulated datasets and the actual LGD and expected LGD calculated for each account. The MSE is calculated for each account and the average per dataset is reported in Figure 19. Each bar on this chart gives the level of the MSE and indicates what portion of the MSE is due to the variance and what portion is due to the squared bias.

The numbering (1 to 5) on the bar graphs in Figure 19 corresponds to the numbering of each simulated dataset as set out in Table 4. Table 4 indicates what parameter estimates are used to simulate these datasets.

Not only does the DWSA model yield the lowest MSE, but all of its components also perform best in that it yields the lowest squared bias and variance on all five simulated datasets. Figure 19 contains the results of the simulated data. Results, when ranked from best to worst performing, rank the same for the simulated data as for the actual data, as discussed in Section 5.1 above. Once again the improvements suggested by this paper culminating in the DWSA model, do indeed yield results more favourable to those achieved by the EWSA model. It therefore holds true that by default weighting

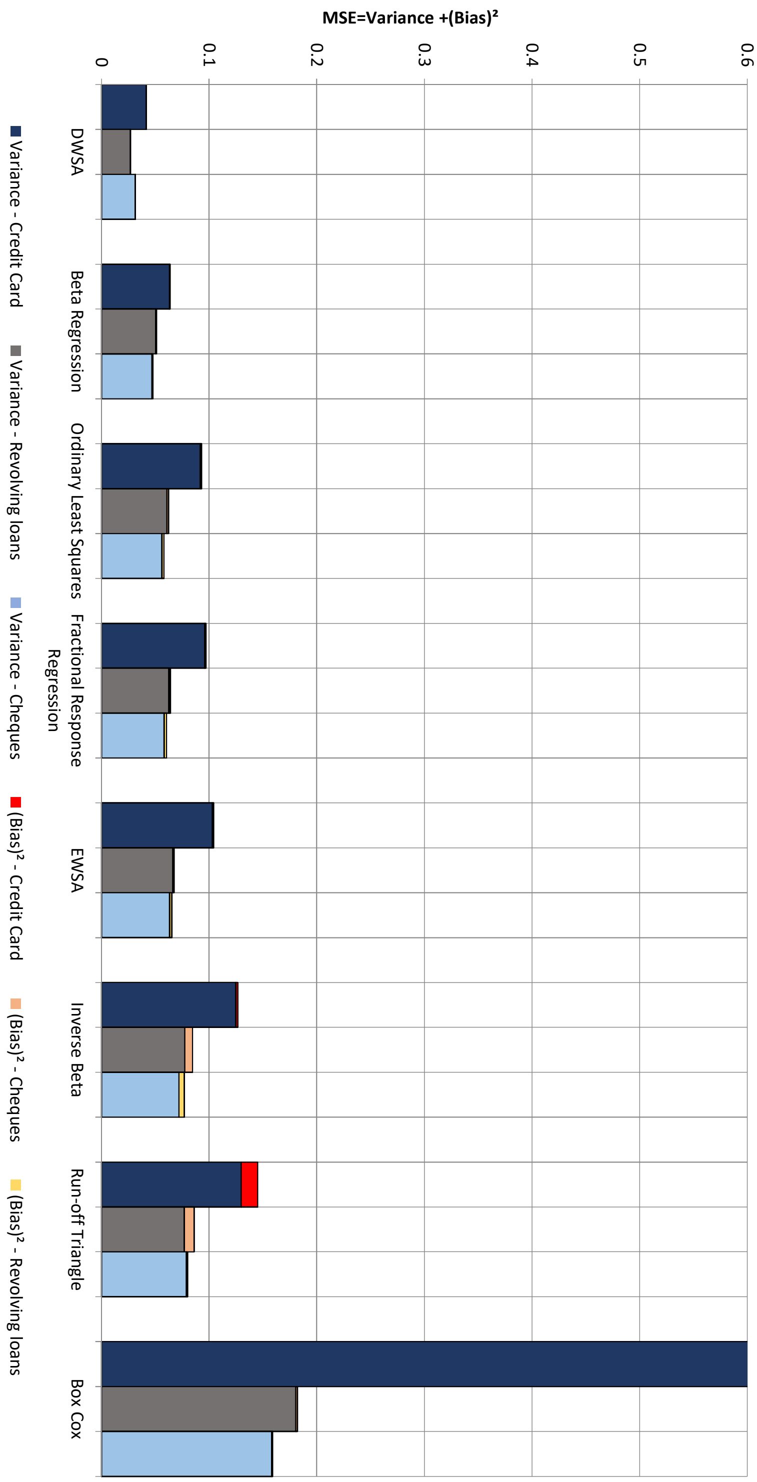


Figure 17. Retail bank dataset results for direct modelling approaches.

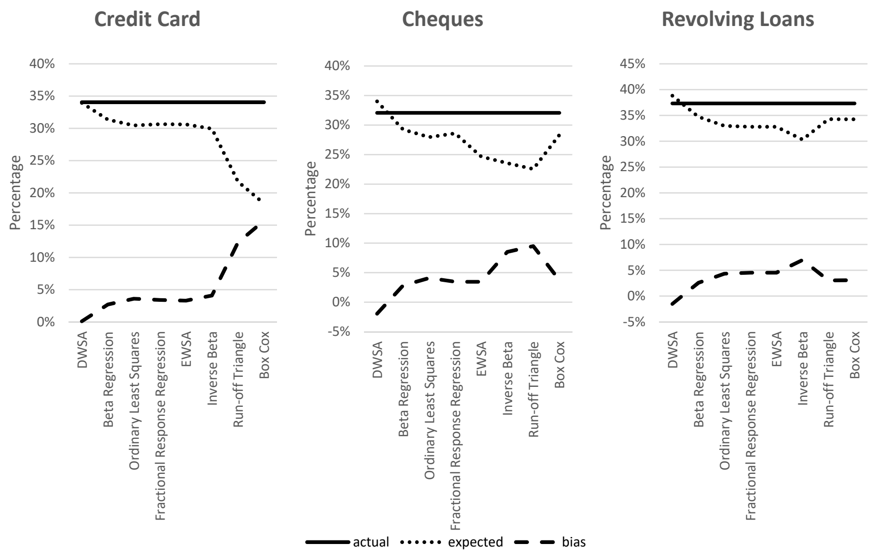


Figure 18. Bias, actual recovery rate and expected recovery rate for the retail bank datasets.

LGD estimates, adding negative cash flows into LGD modelling and by catering for over-recoveries, the model's MSE decreases. As per the retail data, simulated data indicate that the popularly used run-off triangles are outperformed by all other models in this comparison, with one exception: the Box Cox model, which yields the highest MSE. More detail on the biases is displayed graphically in Figure 20.

The bias is smallest in the DWSA model. The DWSA yields the lowest bias on all five simulated datasets. The average bias on the DWSA model is −0.82%. The beta regression model yields an average 2.73% bias. This model performs second best when comparing biases. The other models deliver disappointing bias values. The conclusion for the biases in the DWSA model remains the same as for retail data. The improvements suggested by this paper deliver a superior MSE, bias and variance when simulated data are used.

6. Conclusion

Traditionally there are seven models typically used to model LGD estimates, with varying success. The models are: beta regression, inverse beta model, fractional response regression, ordinary least squares regression, exposure weighted survival analysis (EWSA), run-off triangle and Box–Cox transformation. Improvements introduced by this paper were included to align modelling with regulatory requirements and have lead to the introduction of the default weighted survival analysis (DWSA) methodology. A further enhancement to the existing EWSA modelling technique, as introduced by this paper, was to cater for negative cash flows and over-recoveries, as these events occur in practice.

Retail product data for three different types of products are used in the testing of actual data. Five datasets are simulated to further test the accuracy of the various models. The eight datasets, collectively, are representative of datasets that you would typically use to estimate LGD in a retail environment. MSE, bias and variance on both retail and simulated data across the board are lowest

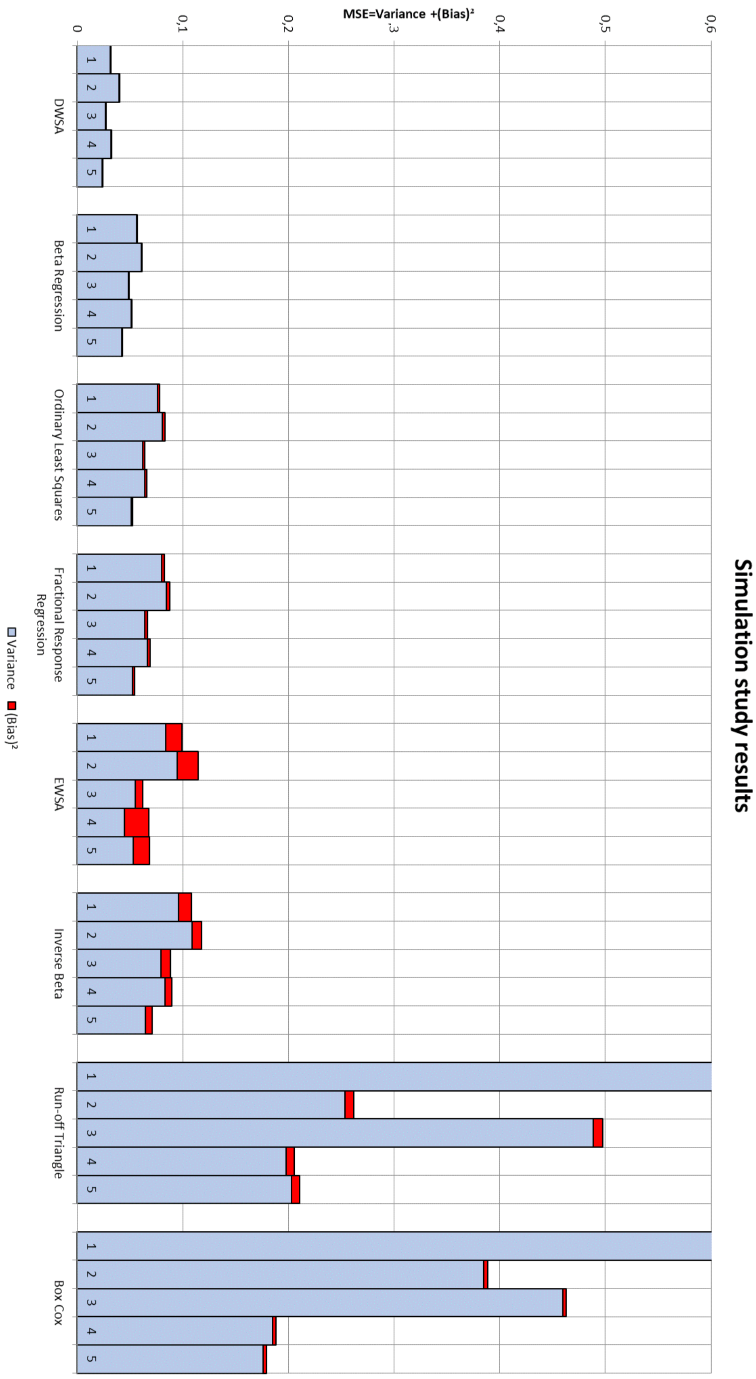


Figure 19. Simulation study results for direct modelling approaches.

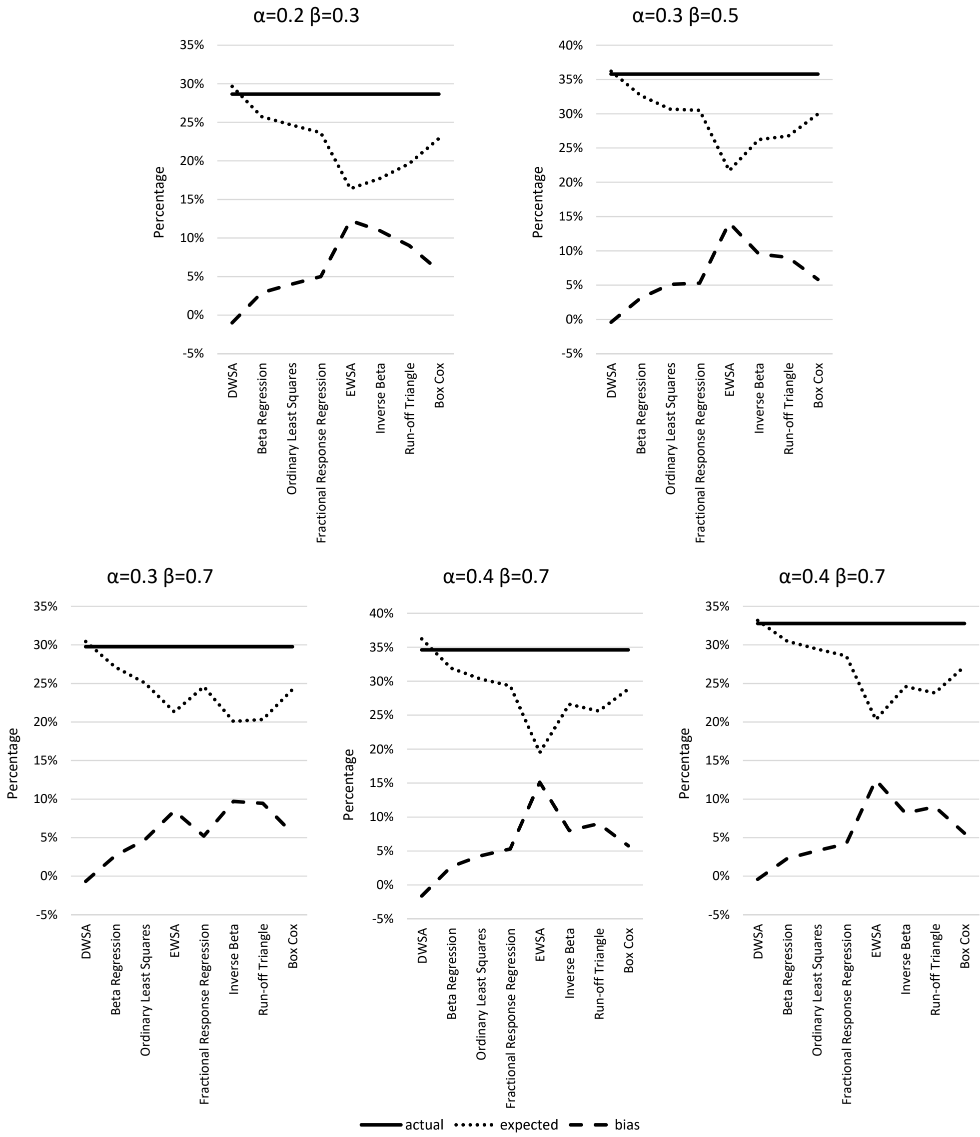


Figure 20. Bias, actual recovery rate and expected recovery rate for the simulated datasets.

for the DWSA model when compared to all other models. The beta regression model performs second best. The run-off triangle method, often used in practice, consistently underperforms most other models.

It is the conclusion of this paper that the improvements suggested firstly serve to introduce a new methodology to estimate LGD. Secondly, as mentioned above, the improvements serve to bring the LGD model in closer alignment to requirements set by regulation. The third contribution by this paper is to improve existing LGD modelling techniques as evidenced by improved MSE, variance and bias.

Similar to how Miu and Ozdemir (2017) adapted Basel LGD modelling techniques to model the IFRS 9 LGD, future research could focus on extending the DWSA method used on Basel models to IFRS 9 models. In addition, the generalised additive proportional hazard model (Hastie and Tibshirani, 1990, pp. 211–218) may be used to allow for time-varying covariates into the DWSA model as the topic of future research. Additional topics of further research can be to use B splines (Ohlsson and Johansson, 2010, pp. 106–108) as the smoothing function for each of these covariates.

Appendix

A.1 Beta regression

Brown (2014, pp. 65–66) suggests making use of a beta regression to model the recovery rate, where LGD is equal to one minus the recovery rate. The beta distribution is reparametrised and covariates are modelled onto the new parameters.

Let the recovery rate be the dependent variable y. The beta density, with parameters ω and τ, is expressed as

$$f(y;\omega;\tau) = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)\Gamma(\tau)} y^{\omega-1}(1-y)^{\tau-1}, \quad 0 \leq y \leq 1, \omega\tau > 0,$$

with

$$\mathrm{E}(Y) = \frac{\omega}{\omega+\tau}$$

and

$$\mathrm{Var}(Y) = \frac{\omega\tau}{(\omega+\tau)^2(\omega+\tau+1)}.$$

The aim is to derive a log-likelihood for a beta regression. Firstly, the above equation is reparametrise to have a location parameter $\mu = E(Y)$ and precision parameter $\phi = \omega + \tau$. Let $\sigma^2 = Var(Y)$. It follows that:

$$\sigma^2 = \frac{\mu(1-\mu)}{(\omega+\tau+1)} = \frac{\mu(1-\mu)}{(\phi+1)}.$$

The initial parameters can now be expressed as a function of the new parameters, $\omega = \mu\phi$ and $\tau = \phi - \mu\phi$. Sub-models for each of the new parameters μ and ϕ will be developed. The sub-model for the location parameter μ is

$$\mu_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}.$$

The sub-model for the precision parameter is $\phi_i = \exp(-w_i\delta)$, where x_i are covariate values for account i and w_i are constant values. A log-likelihood function for the ith observation of the beta regression is given as

$$l(\omega, \tau, y_i) = \ln\Gamma[\omega + \tau] - \ln\Gamma[\omega] - \ln\Gamma[\tau] + [\omega - 1]\ln(y_i) + [\tau - 1]\ln(1 - y_i).$$

A.2 Ordinary least squares

When using linear regression, the LGD is modelled by using the direct modelling approach, where LGD = 1 − recovery rate. The recovery rate is defined as all net cash flows on an account post default and is inclusive of all receipts, fees and costs associated therewith. Actual and predicted recoveries are discounted to the point of default and the LGD is estimated as

$$\text{LGD} = 1 - \frac{\sum \text{present value of observed and predicted future recoveries}}{\text{exposure at default}} = 1 - \text{recovery rate}.$$

The recovery rate is taken as the response variable y and m characteristics describing the loan are taken as covariates, $x_1, x_2, \ldots, x_m$. The linear regression model is given as $y = \mathbf{X}\beta + \varepsilon$.

The model parameters for $\mathbf{b}$ can be retrieved by solving

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y.$$

Advantages of linear regression models are that they are easy to implement, easy to interpret and the parameters are easy to estimate. A disadvantage is that possible non-linear trends in the recovery rate will not be accounted for when using linear regression.

A.3 Fractional response regression

Fractional response regression is used to model the recovery rate and is described in the article written by Bastos (2010, p. 2512). The recovery rate is taken as the dependent variable y with expected value $E(y \mid \mathbf{X}) = G(\mathbf{X}\beta)$ where $0 < G(\mathbf{X}\beta) < 1$. The functional form of $G(\cdot)$ is taken as the logistic function,

$$G(\mathbf{X}\beta) = \frac{1}{1 + \exp(-\mathbf{X}\beta)}.$$

The Bernoulli log-likelihood function

$$l(\beta_i; y_i) = y_i \log(G(x_i\beta_i)) + (1 - y_i)\log(1 - G(x_i\beta_i))$$

is maximised to obtain an estimate for β_i.

A.4 Inverse beta

Brown (2014, p. 64) applies a cumulative beta distribution

$$\beta(y; a; b) = \frac{\Gamma(a + b)}{\Gamma(a + b)} \int_0^y v^{a-1}(1 - v)^{b-1} dv$$

to the recovery rate y where $\Gamma()$ denotes the gamma function, and estimates the parameters

$$a = \frac{\mu^2(1 - \mu)}{\sigma^2} - \mu$$

and

$$b = a\left(\frac{1}{\mu} - 1\right).$$

The inverse standard normal cumulative distribution function is then taken to produce the value

$$y_i^* = N^{-1}\left(\beta\left(y_i; a; b\right)\right).$$

An ordinary least squares regression is applied to y_i^* and the transformation is applied in reverse to get the predicted recovery rate $\widehat{y}_i$.

A.5 Run-off triangles

A run-off triangle contains cells that correspond to accounts defaulting in month i and being k months in default. Each cell contains the cumulative cash flows $C_{i,k}$. The following matrix illustrates a run-off triangle where the end of the workout period is indicated by n. The values for $C_{i,k}$ are observable where $i + k <= n + 1$ and need to be predicted for $C_{i,n}$ with $i = 2, \ldots, n$. The chain ladder approach does this recursively, $\widehat{C}_{i,k} = \widehat{C}_{i,k-1}\widehat{f}_k$ with starting value $\widehat{C}_{i,n+1-i} = \widehat{C}_{i,n+i-1}$ and $\widehat{f}_k = \sum_{i=1}^{n+1-k} C_{i,k}/C_{<,k-1} = (\sum_{i=1}^{n+1-k} C_{i,k-1}/C_{<,k-1})F_{i,k}$ a weighted average of the development factor $F_{i,k} := C_{i,k}/C_{i,k-1}$, where $C_{<,k-1} = \sum_{i=1}^{n+1-k} C_{i,k-1}$ (Braun, 2004, p. 401).

$$\begin{array}{cccc} & 0 & k & n \\ 0 & C_{0,0} & C_{0,k} & C_{0,n} \\ & & & \\ i & C_{i,0} & C_{i,k} & \\ & & & \\ n & C_{n,0} & & \end{array}$$

A.6 Box–Cox transformation

The Box–Cox transformation,

$$\begin{cases} \frac{(y_i+c)^\lambda - 1}{\lambda} & \text{if } \lambda = 0, \\ \log(y_i + c) & \text{if } \lambda \neq 0, \end{cases}$$

is applied to the recovery rate variable y_i and the parameters λ and c are calculated. Ordinary least squares is applied to the transformed variable and the transformation is applied in reverse (Brown, 2014, p. 66).

References

Bastos, J. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance*, **34**, 2510–2517.

BCBS (2006). *International Convergence of Capital Measurement and Capital Standards. A Revised Framework*. Bank for International Settlements, Basel.

Braun, C. (2004). The prediction error of the chain ladder method applied to correlated run-off triangles. *ASTIN Bulletin: The Journal of the IAA*, **34**, 399–423.

Brown, I. L. J. (2014). *Developing Credit Risk Models Using SAS Enterprise Miner and SAS/STAT: Theory and Application*. SAS Institute, Cary, North Carolina.

CHEN, R. AND WANG, Z. (2013). Curve fitting of the corporate recovery rates: The comparison of beta distribution estimation and kernel density estimation. *PLoS ONE*, **8**, 1–9.

COLLET, D. (2003). *Modelling Survival Data in Medical Research*. Second edition. Chapman & Hall / CRC, London.

GREENE, W. H. (2003). *Econometric Analysis*. Fifth edition. Prentice Hall, Upper Saddle River, New Jersey.

HASTIE, T. J. AND TIBSHIRANI, R. J. (1990). Generalized additive models. *In Monographs on Statistics and Applied Probability*, volume 43. Chapman & Hall, London.

JIMENEZ, G. AND MENCIA, J. (2009). Modelling the distribution of credit losses with observable and latent factors. *Journal of Empirical Finance*, **16**, 235–253.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Second edition. Wiley & Sons, Hoboken, New Jersey.

MIU, P. AND OZDEMIR, B. (2017). Adapting the Basel II advanced internal-ratings-based models for International Financial Reporting Standard 9. *Journal of Credit Risk*, **13**, 53–83.

OHLSSON, E. AND JOHANSSON, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, Berlin.

WITZANY, J., RYCHNOVSKY, M., AND CHARAMZA, P. (2012). Survival analysis in LGD modelling. *European Financial and Accounting Journal*, **7**, 6–27.

Manuscript received 2017-10-29, revised 2018-04-03, accepted 2018-04-25.

Chapter 2
Section 3

Errata:

Default Weighted Survival Analysis to Directly model Loss Given Default.

Errata for the paper: ”Default weighted survival analysis to directly model loss given default”.

1. Page 38, Line 11 of Abstract: Change “default weight LGD” to “default weighted LGD”.
2. Page 38, Line 14 of Abstract: Change “the models more are closely” to “the models are more closely”.
3. Page 41, Line 19 from top: Change “survived the default rate” to “survived the default state”.
4. Page 41, Line 8 from bottom: Change “$\lambda(t)$” to “$h(t)$”.
5. Page 42, Line 19 from top:

   The Breslow-Crowley form for the maximum likelihood estimator of the baseline hazard change to

$$\hat{h}_0\left(t\right) = \frac{\sum\limits_{i=1}^{k} dN_i(t)}{\sum\limits_{i=1}^{k} exp(\mathbf{x_i}'\beta)Y_i(t)}.$$

6. Page 45: The formula in the middle of the page given for the Kaplan-Meier estimate of the survival curve:

   $S\left(t\right)$ change to $\hat{S}\left(t\right)$.
7. Page 46, Line 3 from the top: The equation change to:

$$OR = \left| Min\left(Min_i\left(E\hat{A}D_{i,0} - \sum_{s=1}^{T_w} D\hat{C}F_{i,s} \right), 0 \right) \right|.$$

8. Page 46, Line 5 from top: The adjusted survival curve change to

$$S^*\left(t,i\right) = \frac{EAD_{i,0} - \sum\limits_{s=1}^{t} DCF_{i,s} + OR}{EAD_{i,0} + OR}.$$

9. Page 46, Line 15 from the top changes to:

$$MR\left(t,i\right) = MR^*\left(t,i\right) \times R^*\left(t-1,i\right).$$

10. Page 64 Line 13 from the bottom, it is stated that $\omega t > 0$. This changes to ω, $t > 0$.

Chapter 3

Making use of Survival Analysis to indirectly model Loss Given Default.

Chapter 3
Section 1

Guidelines for authors submitting an article to the Operations Research Society of South Africa.

Instructions to Authors

Submission of manuscripts
Anonymous papers (accompanied by a cover e-mail detailing the names and affiliations of authors) may be submitted electronically (preferably as postscript or pdf documents typeset in LaTeX) to the Editor-in-Chief via ORiON's online submission system at `http://www.orssa.org.za/ojs` — the only *other* file format that will be accepted is MS Word documents.

Preparation of manuscripts
Authors are requested to conform to the example paper format available in postscript and pdf formats on the ORSSA webpage `http://www.orssa.org.za` → `ORiON` → `Submissions` → `Example of Paper Format`. This format is also supported by the ORiON LaTeX style sheet (which may be downloaded from `http://www.orssa.org.za` → `ORiON` → `Submissions` → `Style Sheets`). Instructions on how to use these style sheets are available in postscript and pdf formats at `http://www.orssa.org.za` → `ORiON` → `Submissions` → `Instructions for Style Sheets`.

Author and affiliation details
The names of all authors, their affiliations, postal addresses, e-mail addresses and fax numbers should be included in a cover letter or e-mail accompanying submissions. These items will be incorporated into the manuscript *by the journal manager upon acceptance* (submissions should not originally include this information, so as to facilitate blind peer review).

Abstracts and key words
Papers submitted in English should be preceded by an abstract not exceeding approximately 300 words in length. However, all papers not in English should be accompanied by an *extended and detailed* abstract in English of about 1 000 words in length, in addition to a brief abstract in the language of submission (not exceeding approximately 300 words in length). In all cases a list of suitable key words should be listed directly after the abstract, so as to facilitate searches in electronic databases to which ORiON abstracts are contributed.

Mathematical formulae
All mathematical formulae should form part of sentences (and should hence include punctuation, where necessary, but should not be preceded by colons). Mathematical formulae and expressions should be typeset in text lines where possible, the only exceptions being formulae that are so bulky that they would force increased line spacing if included in the text, or formulae that have to be numbered for further referencing.

Formatting
All Latin abbreviations or phrases, such as *e.g., i.e., et al., vice versa, etc.* should be typeset in italics. If MS Word is used to prepare a manuscript, it should be utilised appropriately. For example, **all** mathematical formulae and expressions should be typed in Microsoft Equation Editor (and not merely as italicised text) and section headings should be typeset as *headings* (and not as enlarged, bold faced normal text). Both the full stop and comma are acceptable as decimal separators — however, a choice between these separators should be made and applied consistently by authors.

Figures and tables
Figures and Tables should be numbered consecutively, using separate numbering sequences (*e.g.* Table 1, Table 2, Figure 1, Table 3, Figure 2, . . . rather than Table 1, Table 2, Figure 3, Table 4, Figure 5, . . .). Tables and figures should be accompanied by detailed captions and should be included in the main body of text (not on separate pages at the end of the manuscript). Authors need not include separate high quality photographs or electronic copies of figures when submitting manuscripts — these will be requested by the journal manager (if necessary) upon acceptance of the manuscript. All Figures and tables should be referenced in the text.

Theorems, algorithms and other numbered environments
Theorems, Algorithms and other numbered environments should be numbered consecutively, using separate numbering sequences (*e.g.* Theorem 1, Theorem 2, Algorithm 1, Corollary 1, Algorithm 2, ... rather than Theorem 1, Theorem 2, Algorithm 3, Corollary 4, Algorithm 5, ...). These environments are supported by the official ORiON LaTeX style sheet — further information on how to utilise these environments in LaTeX may be found at `http://www.orssa.org.za` → `ORiON` → `Submissions` → `Instructions for Style Sheets`.

Literature citations
Authors have a choice whether to follow the Harvard (author date) standard or the Vancouver (numerical) standard for literature citations — one of these standards should be applied consistently. Footnotes should not be used for citation purposes. All items in the bibliography should be cited in the text.

According to the Harvard standard literature citations in the text should proceed by listing the relevant author's name and the year of publication (*e.g.* "An optimal solution exists (Dantzig 1963)." or "According to Dantzig (1963) an optimal solution exists."). Additional information, such as page numbers, chapter numbers, theorem numbers, *etc.*, may be given directly after the date, separated by a comma (*e.g.* "An optimal solution exists (Dantzig 1963, p. 69)." or "According to Dantzig (1963, p. 69) an optimal solution exists."). For literature citations involving two authors, both authors' names should be listed, separated by an amprasand (*e.g.* "An optimal solution exists (Dantzig & Wolfson 1967, Theorem 4.2)." or "According to Dantzig & Wolfson (1967, Theorem 4.2) an optimal solution exists."). For literture citations involving more than two authors, only the first author's name should be listed in conjunction with the phrase *et al.* (*e.g.* "An optimal solution exists (Dantzig *et al.* 1972, §3)." or "According to Dantzig *et al.* (1972, §3) an optimal solution exists."). In cases of more than one bibliography entry per author per year, small alphabetical characters should be used to distinguish between references (*e.g.* "An optimal solution exists (Dantzig 1965b)." or "According to Dantzig (1963b) an optimal solution exists.").

According to the Vancouver standard literature citations in the text should proceed by listing the number of the relevant bibliography entry (*e.g.* "An optimal solution exists [7]." or "According to Dantzig [7] an optimal solution exists."). Additional information, such as page numbers, chapter numbers, theorem numbers, *etc.*, may be given directly after the citation number, separated by a comma (*e.g.* "An optimal solution exists [7, p. 69]." or "According to Dantzig [7, p. 69] an optimal solution exists."). For literature citations involving two authors, both authors' names may be listed, separated by an amprasand (*e.g.* "An optimal solution exists [9, Theorem 4.2]." or "According to Dantzig & Wolfson [9, Theorem 4.2] an optimal solution exists."). For literature citations involving more than two authors, only the first author's name may be listed in conjunction with the phrase *et al.* (*e.g.* "An optimal solution exists [10, §3]." or "According to Dantzig *et al.* [10, §3] an optimal solution exists.").

A more comprehensive list of citation examples (using both standards) may be found at `http://www.orssa.org.za` → `ORiON` → `Submissions` → `Example of Paper Format` by clicking on the link `Examples of Reference Citations and Bibliography Listings`.

References
Books should be listed in the bibliography by including the surnames and initials (without punctuation) of all authors and/or editors (IN SMALL CAPITALS), the date of publication, the title (*in italics*, using small letters only, the only exceptions being the first word of the title and proper nouns), the edition (if second or higher), the publisher, the city of publication (followed by the official two-letter abbreviation of the state for cities in the United States — no country names should be listed), and the relevant pages cited (if appropriate), such as in the examples below:

[1] Dantzig B, 1963, *Linear programming and extensions*, 2^{nd} Edition, Princeton University Press, Princeton (NJ).

[2] Gendreau M, Laporte G & Potvin J-Y, 2002, *Metaheuristics for the capacitated vehicle routing problem*, pp. 129–149 in Toth P & Vigo D (Eds.), *The vehicle routing problem*, SIAM, Philadelphia (PA).

Journals should be listed in the bibliography by including the surnames and initials of all authors (in small capitals), the date of the issue, the title of the relevant paper (*in italics*), the title of the journal (not abbreviated), the volume (and issue/part) number (**in bold face**), and the pages of the relevant paper, such as in the example below:

[3] Norese MF & Toso F, 2004, *Group decision and distributed technical support*, International Transactions in Operational Research, **11(4)**, pp. 395–417.

Online resources should be listed in the bibliography by including the surnames and initials of the web page designer (if known, in small capitals), the date of construction of the web page (if known), the title of the web page (if known, *in italics* — this is typically found in the title bar at the very top of the web page), an indication that it is an online reference, the date on which the site was accessed, and the URL (`in true type or courier fonts`), such as in the example below.

[4] Skiena SS, 1997, *The algorithm design manual*, [Online], [Cited September 9th, 2004], Available from `http://www2.toki.or.id/book/algdesignmanual/index.htm`

Theses and dissertations should be listed in the bibliography by including the surnames and initials of the author, the date, the thesis (or dissertation) title, the university where the thesis (or dissertation) was submitted and the city in which the university is situated, such as in the example below [5]. An example of an unpublished technical report [6] is also shown below.

[5] Vumbi AI, 2003, *Algorithmic complexity*, MSc Thesis, University of Stellenbosch, Stellenbosch.

[6] Hamming R, 1956, *On the amount of redundancy required to correct information errors*, (Unpublished) Technical Report TR 1956-371, Bell Laboratories, Murray Hill (NJ).

An example of the format in which an unpublished conference paper should be listed in the bibliography is given in [7] below, whilst an example of the bibliography listing format of a paper published in conference proceedings is shown in [8] below.

[7] Lacomme P, Prins C & Ramdane-Chérif W, 2002, *Fast algorithms for general arc routing problems*, Paper presented at the 16^{th} Triennial Conference of the International Federation of Operations Research Societies, Edinburgh.

[8] Wilkinson C & Gupta SK, 1969, *Allocating promotional effort to competing activities: A dynamic programming approach*, Proceedings of the 5^{th} Triennial Conference of the International Federation of Operations Research Societies, Venice, pp. 419–432.

The bibliography should be arranged in alphabetical order, according to first author surnames.

Note that although authors may use either the Harvard standard or the Vancouver standard (consistently) for citation purposes in the text, all references in the bibliography are expected to adhere to the guidelines above — irrespective of which citation standard is utilised by authors. A more comprehensive list of referencing examples may be found at `http://www.orssa.org.za` → `ORiON` → `Submissions` → `Example of Paper Format` by clicking on the link `Examples of Reference Citations and Bibliography Listings`.

Chapter 3
Section 2

Article Title:

Making use of Survival Analysis to indirectly model Loss Given Default.

Article Authors:

M. Joubert, T. Verster and H. Raubenheimer.

The article was accepted for publication by the Operations Research Society of South Africa (2018).

Contents of Chapter 3 Section 2

List of Figures

List of Tables

Operations Research Society of South Africa

Submitted for publication in ORiON

Operasionele Navorsingsvereniging van Suid-Afrika

Making use of Survival Analysis to indirectly model Loss Given Default

Abstract

A direct or indirect modelling methodology can be used to predict LGD. When using the indirect LGD methodology, two components exist, namely the loss severity component and the probability component. Commonly used models to predict the loss severity and the probability component are the haircut- and the logistic regression models respectively. In this article, survival analysis was proposed as an improvement to the more traditional logistic regression method. The MSE (mean squared error), bias and variance for the two methodologies was compared and it was shown that the improvement enhanced the model's predictive power. The proposed LGD methodology(using survival analysis) was applied on two simulated datasets and two retail bank datasets, and outperformed the logistic regression LGD methodology. Additional benefits included that the new methodology could allow for censoring as well as predicting probabilities over varying outcome periods.

Key words: loss given default, survival analysis, indirect approach, Basel.

1 Introduction and Literature Review

Retail banks, following the internal rating based approach, model their own estimates for probability of default (PD), loss given default (LGD) and exposure at default (EAD). These components are used to calculate regulatory capital. A distinction can be made between subjective and objective LGD methodology. Subjective LGD methodology makes use of expert judgement and are used for low default portfolios, portfolios with insufficient data and new portfolios. Objective LGD methodology can be classified into explicit and implicit methodology. Explicit methodology allows for the direct computation of LGD, whereas with implicit methodology LGD relevant information needs to be extracted by applying applicable procedures. The market LGD, implied market LGD and the workout LGD are categorized as objective LGD methodology and expert judgement is categorized as a subjective method (Engelmann & Rauhmeier, 2011, p.157). The workout LGD is used in the retail sector, and the market LGD and implied market LGD in the corporate sector. The market LGD is calculated as one minus the recovery percentage derived from the corporate bond price or share price available at the point of default. The implied market LGD is modelled from risky but not defaulted corporate bond or shares prices by making use of a theoretical asset pricing model (BCBS, 2005, p.4). The workout LGD can be modelled by using the direct approach or the indirect approach. When using the direct approach, the LGD is equal to one minus the recovery rate (De Jongh et al. 2017, p.261). The indirect approach uses two components that are modelled separately namely the probability component and the loss severity component. The market LGD is an example of an ex-post or actual LGD and the workout LGD is an example of an ex-ante or estimated LGD (Engelmann & Rauhmeier, 2011, p.157-158). Figure 1 contains a diagram that illustrates the classification of the various LGD approaches.

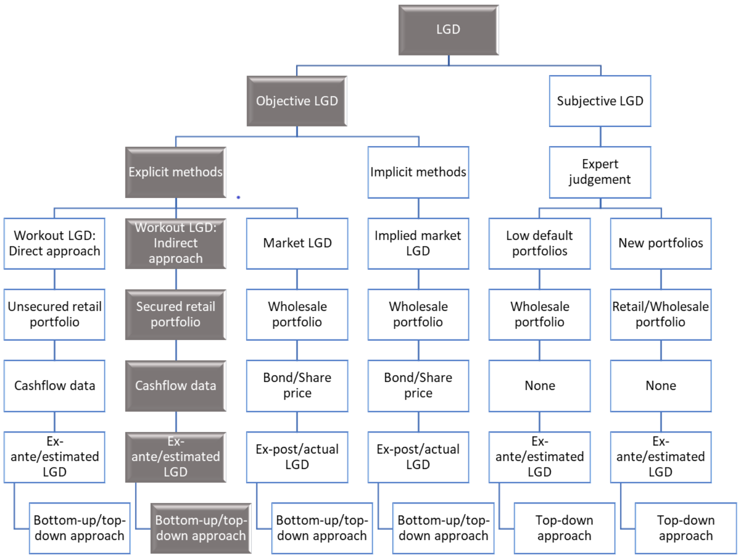


Figure 1: *LGD approaches classified.*

The workout LGD models used in the retail sector is not as advanced as the market LGD or implied market LGD models used in corporate loans due to the fact that most of the work on prediction of LGD pertains to the corporate sector (Qi & Yang, 2009, p.788). Retail data was not stored appropriately in the past and consequently retail methodologies are not well established. Corporate bond prices and share prices are publicly available at the point of default and used to infer the relative credit risk of the underlying company, the associated risk premium and the recovery percentage (Leow & Mues, 2012, p.184). The papers by Lotheram et al. (2012), Qi & Zhao (2011) and Bellotti & Crook (2012) contain comparisons between different LGD modelling techniques. This article will focus on the ex-ante indirect workout LGD used in the secured retail sector and the term will be simplified to LGD. This approach is highlighted in Figure 1.

For several decades, the focus in the retail sector was the probability of default model (Tong, Mues & Thomas, 2013, p.548). PD models have been used in the retail sector since 1960 (Anderson, 2007, Section xxiv). With the advent of regulatory capital changes driven by legislation such as the Basel Capital Accord (BCBS, 2006), more focus and emphasis has been placed on LGD methodology. A top-down approach or bottom-up approach can be followed to model LGD. The average LGD per segment is calculated for the top-down approach. Account-level estimates are estimated by making use of regression techniques when using the bottom-up approach (Engelmann & Rauhmeier, 2011, p.158).

Zhang & Thomas (2012), Schmidt (2006) and Witzany et al. (2012) respectively made use of a linear regression, run-off triangle and Cox proportional hazards regression to model the recovery rate directly. Tong et al. (2013) made use of a zero adjusted gamma model to model the LGD directly. The LGD was modeled indirectly by Somers & Whittaker (2007), Qi & Yang (2009), Zhang et al (2010), Tong et al (2011) and Leow & Mues (2012).

Leow & Mues (2012) indirectly modelled LGD by modelling a probability component and loss severity component and combining them to estimate LGD. The probability component is modelled by making use of a logistic regression with binary outcomes: write-off or not write-off. Incomplete accounts are grouped as not written-off accounts. The European Banking Authority (2017, p.34) mentions that incomplete accounts carry valuable information and excluding this information will lead to the underestimation of LGD estimates.

Survival analysis instead of logistic regression will be used to model the probability component of the indirect LGD model. The advantages of using survival analysis is that incomplete accounts are modelled separately and the need to make assumptions with regards to incomplete accounts are eliminated. Further advantageous of the survival analysis method are that a time-varying outcome period is used, more than two outcomes are allowed for in the target variable of the probability components and survival time are incorporated into the model.

The mathematical notation that is used in this article is provided in Section 2. In Section 3, LGD for the indirect approach is described. The probability component and the loss severity component are described. Section 4 describes the data, Section 5 describes the results and Section 6 concludes.

2 Mathematical Notation

LGD is estimated on accounts that are in default, where default is assigned per the Basel default definition. The Basel default definition stipulates that an account enters into default when the bank considers that the obligor is unlikely to pay or if the obligor is past due by more than 90 days on any material credit obligation (BCBS, 2005).

The loss given default on account i, which has been in default for t months, is estimated as

$$LGD_{i,t} = \frac{E(L_{i,t}|default)}{E_{i,t}},$$

where $E(L_{i,t}|default)$ is the expected loss amount for a defaulted account and $E_{i,t}$ the exposure t months after default.

An account that is in default can either be written off, cure or remain incomplete:

- An account is written-off when the expectation is that there will not be any significant recovery on the outstanding loan amount owed.
- An account in default cures when the default flag is lifted, when the arrears amount is settled and the account becomes up-to-date.

- An account that remains in default at the end of the workout period is deemed to be incomplete.

Define τ as time to the first of:

- exiting the default state, or
- the end of the reference period.

In the case of an incomplete account, the value of τ will be equal to the end of the workout period. The end of the workout period, T_w, is defined as the maximum time that is allowed to recover on an account. Furthermore, let W_τ, C_τ and I_τ be the events of write-off, cure or incomplete at point τ respectively. An account is deemed to be worked out when there are no further recoveries on the account and/or the collection process on the account is complete.

The probability that account i, which is t months in default, will write-off, cure or remain incomplete in the interval $[t, \tau]$ is given by $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$ where $P_{i,t}(W_\tau) + P_{i,t}(C_\tau) + P_{i,t}(I_\tau) = 1$ for every t. The loss amount given write-off, cure or incomplete is given by $L_{i,t}|W_\tau, L_{i,t}|C_\tau$ and $L_{i,t}|I_\tau$ respectively.

The expected value of the loss amount can be written as the sum product of the loss amount components and the probability components. The loss given default for account i, which is t months in default, is then given as

$$LGD_{i,t} = \frac{E(L_{i,t}|default)}{E_{i,t}} = \frac{L_{i,t}|W_\tau \times P_{i,t}(W_\tau) + L_{i,t}|C_\tau \times P_{i,t}(C_\tau) + L_{i,t}|I_\tau \times P_{i,t}(I_\tau)}{E_{i,t}}.$$

Leow & Mues (2012) predict the probability components, $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$, by making use of logistic regression, and they use a haircut model to predict the loss severity components, $L_{i,t}|W_\tau, L_{i,t}|C_\tau$ and $L_{i,t}|I_\tau$. In the modelling methodology section that follows, the logistic regression approach will be replaced by survival analysis. The survival analysis approach has the advantage of using survival time. The logistic model considers a fixed outcome and ignores survival time and censoring. Further advantages and reasons for replacing the logistic regression with survival analysis will be given when concluding in Section 6. The modelling methodology section contains a description of survival analysis and how it is used to predict the probability components, as well as a description of the haircut model used by Leow & Mues (2012) and how the haircut model is used to predict the loss severity component.

3 Modelling Methodology

Leow & Mues (2012) did not differentiate between incomplete accounts and cured accounts as described in Section 2, but combined these two states and predicted either one of the following binary outcomes: write-off or not write-off. Leow & Mues (2012) modelled the probability of write-off, $P_{i,t}(W_\tau)$, by making use of logistic regression

$$\ln\left(\frac{P_{i,t}(W_\tau)}{1-\ P_{i,t}(W_\tau)}\right) = \mathbf{x}_i'\beta,$$

with $\mathbf{x}_i$ a column vector of covariates for account i. For the purpose of this article we will distinguish between incomplete and cured accounts and model write-off, cured and incomplete states.

In this section, we discuss how survival analysis instead of logistic regression can be used to predict the probability components $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$. This is followed by a description of the haircut model that we will use. No changes are made to the haircut model that Leow & Mues (2012) used. The haircut model is one of the components used to calculate LGD and will be described for the sake of completeness.

3.1 Modelling Methodology for the Probability Components

The aim of the probability model is to estimate the probability that an account i, which is t months in default, will write-off, cure or remain incomplete in the interval $[t, \tau]$. These probabilities are given as $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$, where W_τ, C_τ and I_τ are the events of write-off, cure or incomplete at point τ. Survival analysis is used to predict these probability components.

The survival function is defined as the probability of an event occurring after a specified time, t. Witzany et al. (2012, pp. 13) defines the survival function, S, as

$$S(t) = 1 - F(t) = 1 - P(T < t),$$

where the random variable T denotes the time of the event and the cumulative distribution function is denoted as $F(t)$. The corresponding probability density function is $f(t)$. The hazard rate, $h(t) = \frac{f(t)}{S(t)}$, is the instantaneous rate of exit at t, given that survival has been attained up to point t. The survival function, $S(t) = e^{-H(t)}$, is expressed in terms of the cumulative hazard function $H(t) = \int_0^t \lambda(t)\, ds$ (Witzany et al., 2012, pp. 13–14). Survival analysis is traditionally used for analysing the expected duration of time until one or more events happen. An example of an event can be death. In our context, the survival function, $S(t)$, will be defined as the probability that an account that is in default at time t remains in default until the end of the workout period, T_w. An account can exit the default state by either writing-off or curing. An account that remains in default is flagged as incomplete. The probability that the account exits default in the time interval $(t, t + \Delta t]$, given that the account is still in default at t, is $h(t)\Delta t$.

A survival function for write-off, $S^w(t)$, and a survival function for cure, $S^c(t)$, is defined in the following paragraphs. This is followed by a description of the cumulative incidence function. The cumulative incidence function and the fact that

$$P_{i,t}(W_\tau) +\ P_{i,t}(C_\tau) +\ P_{i,t}(I_\tau) = 1$$

is used to combine the two survival functions $S^w(t)$ and $S^c(t)$ to produce the three probabilities $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$.

$S^w(t)$ is defined as the probability that an account that is in default at time t will not write-off before the end of the workout period, T_w. The survival function $S^w(t) = P(T > t)$, with not write-off as the event. Similarly, $S^c(t)$ is defined as the probability that an account that is in default at time t will not cure before the end of the workout period, T_w. The survival function $S^c(t) = P(T > t)$, with not curing as the event.

The two survival functions, $S^w(t)$ and $S^c(t)$, can either be estimated for the entire population or on segments of the population. The Cox proportional hazards model is used to model these survival curves for different segments (Kleinbaum & Klein, 2012, p.120-123).

The general form of the Cox proportional hazards model can be written in terms of survival curves,

$$S(t) = [S_0(t)]^{exp(\mathbf{x}'\beta)}.$$

The formula states that the survival curve at time t is a function of two quantities. The first of these, $S_0(t)$, is called the baseline survival function whilst the second of these is the exponential expression to the linear sum of the covariates. The baseline survival curve, $S_0(t)$, is estimated by selecting a specific segment and calculating the Kaplan–Meier estimate for that segment. The Kaplan–Meier estimate is the empirical survival curve estimated from the data. If the values of an individual's covariate value falls outside the baseline group, the baseline survival curve, $S_0(t)$, will be adjusted,

$$S_0(t)^{exp(\mathbf{x}'\beta)},$$

to yield a survival curve, $S(t)$, that is the estimate for the segments associated with the new covariate values.

The Cox proportional hazards model can also be defined in terms of hazard functions (Witzany et al., 2012, pp. 14–16) as

$$h(t,x) = h_0(t)\, exp(\mathbf{x}'\beta),$$

with the 0 indicating the baseline in the baseline hazard, $h_0(t)$. The baseline hazard is independent of the covariate values, $\mathbf{x}$. The matching survival function is

$$S(t,\mathbf{x}) = exp\left(-\int_0^t h_0(s) exp\left(\mathbf{x}'\beta\right)\right) = S_0(t)^{exp(\mathbf{x}'\beta)},$$

where $S_0(t) = exp\left(-\int_0^t h_0(s)\right)$. The partial likelihood is used to solve for the parameter estimates, β. The partial likelihood for a specific account i, that exits at time t, is defined as

$$L_i(\beta) = \frac{h(t,\ x_{\mathbf{i}})}{\sum_{j\varepsilon A_i} h(t,\ \mathbf{x_j})} = \frac{exp(-\mathbf{x_i}'\beta)}{\sum_{j\varepsilon A_i} exp(-\mathbf{x_j}'\beta)}$$

with $\mathbf{x_i}$ the set of covariates at the point of exiting default and A_i the set of objects in default at t. It is assumed that there is only one exit at time t. Given that there are K accounts, the equation

$$ln(L) = \sum_{i=1}^{K} ln(L_i)$$

is maximized by using the Newton Raphson algorithm to obtain the beta values, β. When modelling LGD, multiple exits may occur and the partial likelihood is adapted to handle ties. An approximation of the partial likelihood is used to solve the parameter estimates in the case where ties occur. The baseline hazard function is assumed to be constant for each unit time interval and are estimated separately. The likelihood function

$$L_t = \prod_{i=1}^{n} \left[h_0\,(t)\,exp(\mathbf{x_i'}\beta)\right]^{dN_i(t)} exp(-h_0\,(t)\,exp(\mathbf{x_i'}\beta)Y_i(t))$$

is then maximized. The indicator $Y_i\,(t)$ indicates that observation i has not exited default at $t-1$ and is incomplete. The indicator $dN_i(t)$ indicates that observation i exited from default at $(t-1, t]$ by curing or writing off. Witzany et al. (2012, pp. 16) gives the Breslow-Crowley form for the maximum likelihood estimator of the baseline hazard as

$$\hat{h}_0\,(t) = \frac{\sum\limits_{i=1}^{n} dN_i(t)}{\sum\limits_{i=1}^{n} exp(\mathbf{x_i'}\beta)Y_i(t)}.$$

The cumulative incidence function is a recursive formula used to convert the survival curves to probabilities and is defined/constructed as follows:

- Define the probability that account i, which is t months in default, will write-off, cure or remain incomplete, in the interval $[t, t+1]$ as $P_{i,t}\,(W_{t+1})$, $P_{i,t}\,(C_{t+1})$ and $P_{i,t}\,(I_{t+1})$ respectively. The initial values, where $t = 0$, for the probabilities are $P_{i,0}\,(I_1) = 1, P_{i,0}\,(W_1) = 0$ and $P_{i,0}\,(C_1) = 0$, since all accounts will be incomplete at the initial default point.
- The recursive formulas with a starting value for $t = 0$ is
  $P_{i,t}\,(C_{t+1}) = P_{i,t}\,(I_{t+1}) \times (1 - \frac{S^c(t+1)}{S^c(t)})$,
  $P_{i,t}\,(W_{t+1}) = P_{i,t}\,(I_{t+1}) \times (1 - \frac{S^w(t+1)}{S^w(t)})$,
  $P_{i,t+1}\,(I_{t+2}) = P_{i,t}\,(I_{t+1}) - P_{i,t}\,(C_{t+1}) - P_{i,t}\,(W_{t+1})$.

A description of the above-mentioned recursive formulas follows. The entire population is initially incomplete, $P_{i,0}\,(I_1) = 1$. The value $(1 - \frac{S^c(t+1)}{S^c(t)})$ represents the percentage of the incomplete accounts that exit default by curing and $(1 - \frac{S^w(t+1)}{S^w(t)})$ represents the percentage of the incomplete accounts that exit default by writing-off. The percentage writing-off and curing is subtracted from the initial incomplete population.

Next, the sums of the one-month probabilities are taken to achieve the probabilities over the interval $[t, \tau]$

- $P_{i,t}\,(C_\tau) = \sum\limits_{k=t}^{\tau} P_{i,k}\,(C_{k+1})$,
- $P_{i,t}\,(W_\tau) = \sum\limits_{k=t}^{\tau} P_{i,k}\,(W_{k+1})$,

- $P_{i,t}\left(I_\tau\right) = \sum_{k=t}^{\tau} P_{i,k}\left(I_{k+1}\right)$.

3.2 Advantages of using Survival Analysis

The advantages of using the Cox proportional hazards approach instead of a logistic model (Kleinbaum & Klein, 2012, p.110-112) include:

- The Cox proportional hazards model predicts probabilities over varying outcome periods; a logistic regression model predicts probabilities over a fixed outcome period.
- The Cox proportional hazards model allows for censoring.
- The Cox proportional hazards model is robust in the sense that the non-parametric estimate of the baseline hazard will closely approximate the parametric baseline hazard. There is typically uncertainty about the form of the parametric model. The non-parametric nature of the Cox proportional hazards model is the safer option.
- The general form of a survival function for the Cox proportional hazards model is given in Section 3 as
$$S\left(t\right) = \left[S_0(t)\right]^{exp(\mathbf{x}'\beta)}.$$
The hazard rate is easily derived from the survival function and is given as
$$h\left(t\right) = h_0(t)exp(\mathbf{x}'\beta).$$
- The exponential expression will ensure non-negative hazard rate estimates. The non-negative hazard rate estimates are a necessity since hazard rates, by definition, should vary between zero and infinity.
- The β values can be estimated, even though the baseline hazard, $h_0(t)$, is unspecified. Once the β values are estimated, the effect of the explanatory variables can be measured through the hazard rate and there is no need to estimate the baseline hazard.

3.3 Modelling Methodology for the Loss Severity Component

In this section, estimation of the loss amount given write-off, cure or incomplete, given by $L_{i,t}|W, L_{i,t}|C$ and $L_{i,t}|I$ are described.

The loss amount given write-off, $L_{i,t}|W$, is modelled using the haircut model as used by Leow & Mues (2012). The loss amount given cure, $L_{i,t}|C$, is assumed to be zero, which is a reasonable assumption as nothing is lost when the account cures. The loss amount given incomplete, $L_{i,t}|I$, is handled with an adjustment to the LGD estimate, as this amount is expected to be very low given that the workout period is chosen such that most accounts are written-off or cured before the end of the workout period.

In the haircut model proposed by Leow & Mues (2012), a loss is incurred when the expected haircut, $h_{i,t}$, is smaller than the loan-to-value ratio at the default point, $LTV_{i,\,0}$, where

$$LTV_{i,\,0} = \frac{M_{i,0}}{V_{i,0}}$$

and the expected haircut

$$h_{i,t} = \frac{P_{i,t}}{V_{i,t}}$$

with $M_{i,0}$ the outstanding loan amount of the asset at the default point, $V_{i,t}$ the valuation of the asset at time t in default and $P_{i,t}$ the net proceeds of the loan at point t in default. Each cashflow component of the net proceeds calculation is summarised in the schematic below:

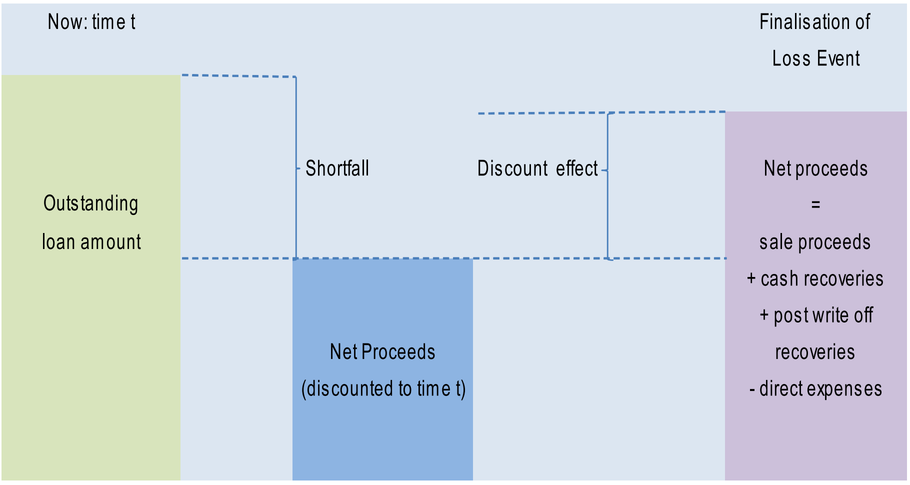


Figure 2: *Net proceeds.*

In Figure 2, sale proceeds represent the realised sale amount of the underlying asset. Cash recoveries represent any additional recovery amounts from observation date through to final realisation of all worked out proceeds, until date of write-off or cure. Post write-off recoveries represent recovery or additional expense amounts post the write-off date. Direct expenses represent legal fees, etc. incurred in realisation of the underlying asset. The net proceeds, $P_{i,t}$, can mathematically be represented as

$$P_{i,t} = (b_{i,t-1} - b_{i,t}) + r_{i,t} - w_{i,t} + g_{i,t}$$

with $b_{i,t}$ the sum of the account balance and the accrued interest. The write-off amount is indicated by $w_{i,t}$ and the value of recoveries made past the write-off point is indicated by $g_{i,t}$. The matched debit interest amount is indicated by $r_{i,t}$.

All net proceeds are discounted back to time t in default. The methodology assumes that the haircut level based on the timing of historical cashflows will be representative of future experience. As mentioned, a loss or shortfall is incurred when the haircut is smaller than the loan-to-value ratio, i.e. $h_{i,t} < LTV_{i,0}$. The shortfall percentage can be defined as

$$Shortfall\,percentage = \frac{M_{i,0}}{V_{i,0}} - \frac{P_{i,t}}{V_{i,t}}.$$

The loss amount given write-off, $L_{i,t}|W$, is then expressed as the expected shortfall percentage given write-off multiplied by the expected valuation of the asset at time t.

$$\begin{aligned} L_{i,t}|W &= E(shortfall\,percentage|W) \times E(V_{i,t}) \\ &= E\left(\frac{M_{i,0}}{V_{i,0}} - \frac{P_{i,t}}{V_{i,t}}|W\right) \times E(V_{i,t}) \\ &= E(LTV_{i,0} - h_{i,t}|W) \times E(V_{i,t}) \\ &= \int_{-\infty}^{LTV_{i,0}} p\,(h)\,(LTV_{i,0} - h_{i,t}\,)dh \times E(V_{i,t}) \end{aligned}$$

where $p(.)$ denotes the probability density function of the distribution for h.

As long as the predicted haircut (ratio of expected discounted sale proceeds and direct costs to the valuation of the asset) exceeds the current defaulted LTV, the net proceeds from the sale will be able to cover the outstanding balance on the loan, i.e. there will be no shortfall. Hence, the expected shortfall, expressed as a proportion of the valuation of the asset is given by the above equation. Leow & Mues (2012, p.191) assume that h follows a normal distribution. This assumption is shown to be realistic in Section 4 below (see Figures 6 and 7 below). Leow & Mues (2012, p.191) make the following transformations

$$z = \frac{h_{i,t} - E\,(h_{i,t})}{\sigma} \sim N\,(0,1)$$

and

$$D = \frac{LTV_{i,0} - E(h_{i,t})}{\sigma}$$

with $h_{i,t}$ the haircut, $LTV_{i,0}$ the default loan to value, $E(h_{i,t})$ the expected value from the haircut and σ the standard deviation of the haircut. The value D is calculated by subtracting the expected value of the haircut from the default loan to value and then dividing it by the standard deviation of the haircut. Hence, the expected shortfall percentage can be expressed as

$$\begin{aligned} L_{i,t}|W &= (\int_{-\infty}^{D} p\,(z)\,(D - z)\sigma dz) \times E(V_{i,t}) \\ &= \{\sigma D \int_{-\infty}^{D} p\,(z)dz - \sigma \int_{-\infty}^{D} p\,(z)\,zdz\} \times E(V_{i,t}) \\ &= \{\sigma D \times \Phi_z\,(D) - \sigma(-\phi_z(D))\} \times E(V_{i,t}) \end{aligned}$$

where $\Phi_z\,(D)$ and $\phi_z(D)$ represent the cumulative distribution function and probability density function of the standard normal distribution (Leow & Mues, 2012, p.191).

4 Data

The two datasets utilised in this paper are obtained from one of the big South African secured retail bank portfolios. These datasets are described in Section 4.1. In addition, two dataset are simulated, which are described in Section 4.2.

4.1 Retail Bank Data

Two secured portfolios are considered; a vehicle and asset portfolio and a home loans portfolio. All source data is extracted from September 2005 until April 2013. From the source dataset, the following fields are either derived or extracted to create the development datasets:

- Entry and exit point (expressed by time in default) for each account in relation to its entry and exit from default.
- Status of the account (cure, loss or incomplete) on exit from default.
- Covariates considered when fitting the cure and loss event models.

An account is deemed to have exited the workout period on the first occurrence of any of the following events:

- Write-off event: When the account is written off or when the legal status on the account indicates insolvency of the individual/juristic person or sale of the underlying asset. Insolvency and sold states are deemed absorbing and the default event deemed completed on occurrence of any of the aforementioned events.
- Cure event: Any account where the reference default flag is lifted, is deemed cured.

The loss given default for the vehicle and asset portfolio and the home loans portfolio of a retail bank is given in Figure 3. The LGD axis in all Figures are left out due to confidentiality.

The development reference period is from May 2011 until April 2013. There are 55 794 accounts on the vehicle and asset development dataset and 66 495 on the home loans development dataset. Macroeconomic, behavioural, application, customer and geographical covariates are included into these models.

The distribution for the vehicle and asset finance LGD is given in Figure 4. Figure 5 contains the LGD distribution for the home loans portfolio. The average vehicle and asset finance LGD is 30.31% and the majority of LGD values are distributed between 6% and 51%. The average home loans portfolio LGD is 8.4% with the bulk of the accounts distributed between LGD values 0% and 18%. The home loans portfolio generally has lower LGD values than the vehicle and asset finance portfolio.

The observed vehicle and asset finance haircut distribution is displayed as the histogram in Figure 6. The mean observed haircut value is $0,706$ and the standard deviation of the

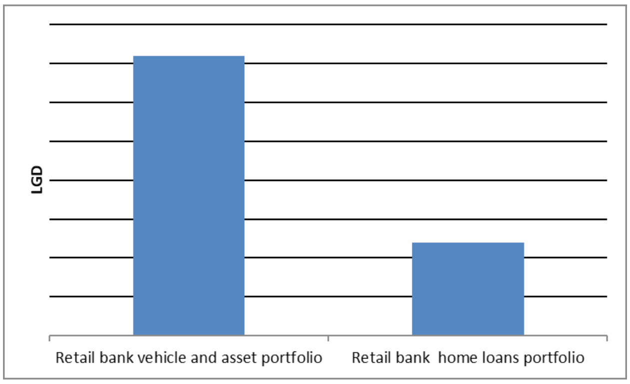


Figure 3: *Retail bank data loss given default.*

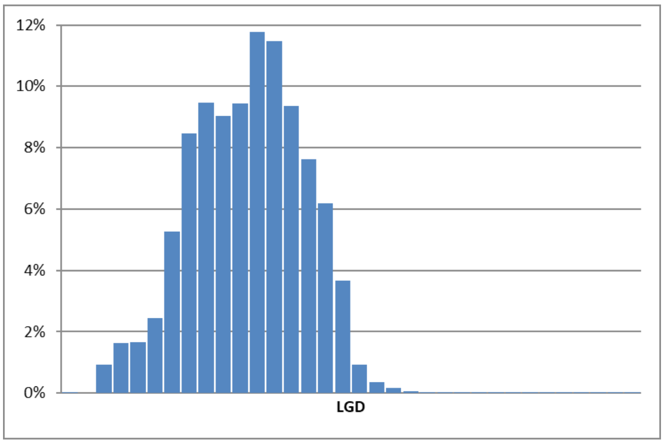


Figure 4: *Vehicle and asset finance LGD distribution.*

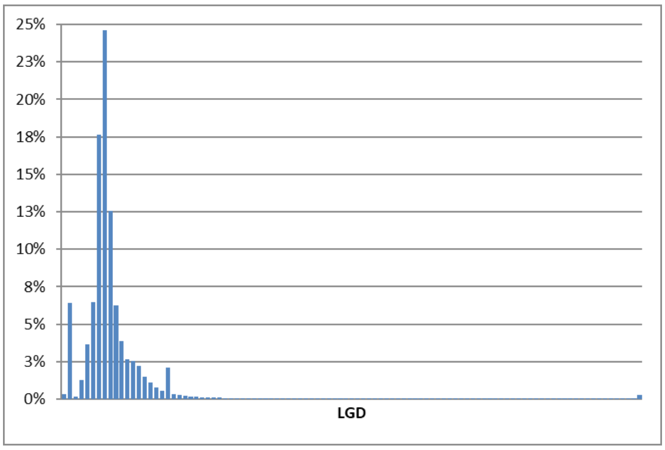


Figure 5: *Home loans LGD distribution.*

observed haircut is $0,228$. A normal distribution curve with the same mean and standard deviation is plotted on the same graphs. One can conclude that the observed haircut for the vehicle and asset finance portfolio follows the normal curve closely. Figure 7 displays the observed home loans haircut distribution with a mean haircut value of $0,428$ and a standard deviation of $0,17$. A normal distribution curve, with the same mean and standard deviation as the actual home loans haircut, is plotted onto Figure 7. The observed home loans haircut follows the normal curve closely. The results for the test for normality in Table 1 confirms that the actual haircut for the home loans and for the vehicle and asset finance portfolio are normal. In both cases the p values are high providing evidences not to reject the null hypothesis that the variable is normally distributed.

		Home Loans	Vehicle and asset finance
Test		p value	p value
Shapiro-Wilk	P <W	0,2168	0,2183
Kolmogorov-Smirnov	P >D	0,1458	>0.1500
Cramer-von Mises	P >W-Sq	0,2358	0,2491
Anderson-Darling	P >A-Sq	0,2221	0,222

Table 1: *home loans and vehicle and asset finance test for normality.*

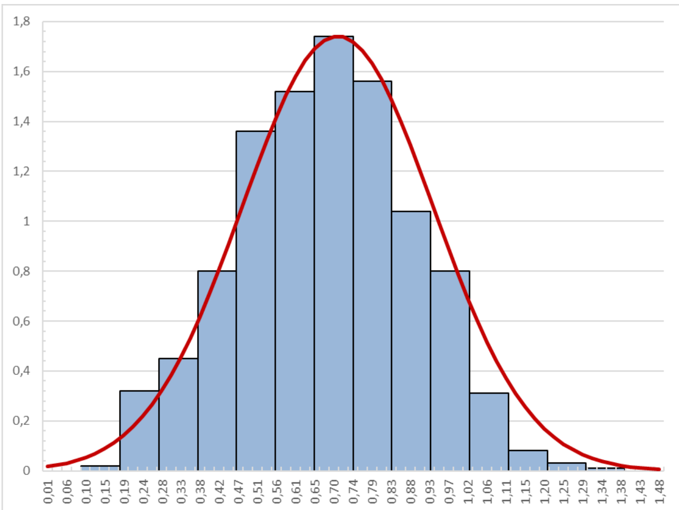


Figure 6: *Vehicle and asset finance haircut distribution.*

4.1.1 Censoring

Censoring is a peculiar feature of survival analysis and occurs when information is known over a certain interval (Klein & Moeschberger, 2003, p.63). There are several reasons why information is not available over the remaining intervals; the event will only occur after the end of the workout period, the event occurs before the observation period started or the information for the account is only available sometime after the first point of default.

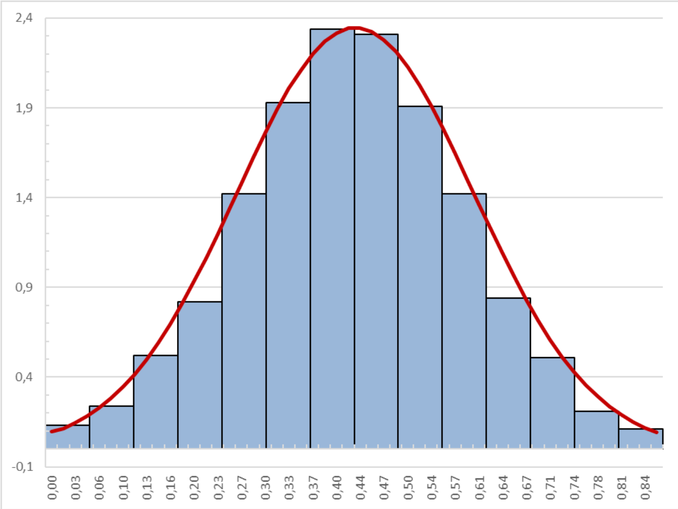


Figure 7: *Home loans haircut distribution.*

Figure 8 indicates the reference period data used for development and is used to describe censoring.

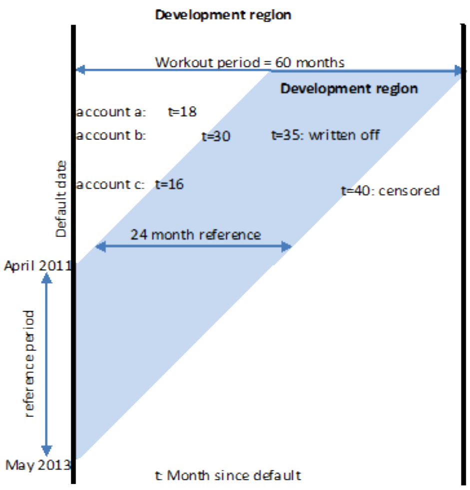


Figure 8: *Development reference period data.*

In Figure 8, account *a* defaults in February 2009, and exits the default status 18 months later due to repossession of the underlying asset. Account *b* defaults in March 2009, and exits the default status due to write-off 35 months later. Account *c* defaults in February 2010, and reaches the end of the workout period without experiencing any event of interest. All these are example of censoring. Censoring is a defining feature of survival analysis and makes survival analysis distinct from other kinds of analysis. The following types of censoring are explained below:

- Left censoring: the event of interest occurred before the observation period started, e.g. account a defaults at time $t = 18$ (Klein & Moeschberger, 2003, p.70).
- Left truncation: the information for an account is only available sometime after the point of default, e.g. account b only has information available after $t = 30$ (Klein & Moeschberger 2003, p.72).
- Right censoring: an account has reached the end of the workout period without experiencing the event of interest, e.g. account c is censored at $t = 40$ (Klein & Moeschberger, 2003, p.64).

Left censoring, right censoring and left truncation are used in the development of the probability of cure and probability of write-off models.

4.2 Simulated Data

The LGD datasets were simulated using the indirect model approach. The simulated datasets are based on the assumed distributions of the actual datasets. However different parameter values are selected to cover a wider range of portfolios than the actual data, since the actual data only represents one possibility for a specific set of parameters. Since the LGD is calculated by combining the probability component and the loss severity component, each of these are simulated separately. These two components were discussed in Section 3 above. Since the three probabilities $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$ are estimated using the survival curves we simulate from predefined survival curves as discussed below. The loss severity component is simulated from the appropriate distribution.

Each of these components were simulated separately. Two survival curves, $S^w(t)$ and $S^c(t)$, were defined in Section 3.1 and combined to produce the three probabilities $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$. Section 3.3 gave an overview of the loss severity component in the form of the expected loss percentage.

The survival time and censoring time for $S^w(t)$ and $S^c(t)$ are simulated for this paper. The article by Bender et al. (2005) describes how to derive the formula that is used to simulate the survival time. The derivation of an equation that is used to simulate the survival time follows:

The Cox proportional hazards model is defined as

$$S(t) = [S_0(t)]^{e^{\mathbf{x}'\beta}}.$$

The baseline survival curve, $S_0(t)$ can be expressed in term of the cumulative hazards rate

$$S_0(t) = e^{-H_0(t)},$$

where the cumulative hazard rate is taken as

$$H_0(t) = \sum_{l=0}^{t} h_0(l).$$

The hazard rate, $h_0(t)$, represents the rate at which objects that have survived until time t, exits at time t. The survival function of the Cox proportional hazards model, $S(t)$, is written as

$$S(t) = [e^{-H_0(t)}]^{e^{\mathbf{x}'\beta}}.$$

The survival function is written in terms of the cumulative distribution function,

$$S(t) = P(T > t) = 1 - F(t),$$

and it follows that

$$F(t) = 1 - \left[e^{-H_0(t)}\right]^{e^{\mathbf{x}'\beta}} = 1 - e^{-H_0(t)e^{\mathbf{x}'\beta}}.$$

Next, let Y be a random variable with distribution function F. It follows that $U = F(Y)$ is uniformly distributed on the interval $[0,1]$, abbreviated as $U \sim Uni[0,1]$. If $U \sim Uni[0,1]$, then $1 - U \sim Uni[0,1]$. Let T be the survival time in the Cox proportional hazards model, it follows that

$$U = e^{-H_0(T)e^{\mathbf{x}'\beta}} \sim Uni[0,1].$$

The inverse of H_0 can be taken when the hazard rate, $h_0(t)$, is positive for all values of t and the random variable, T, in the Cox proportional hazards model can be expressed as

$$T = H_0^{-1}\left[-\log(U)\, exp(-\mathbf{x}'\beta)\right]$$

and U is a random variable with,

$$U \sim Uni[0,1]$$

and $-\log(U)$ is exponentially distributed with parameter 1. The inverse of the cumulative hazard function is given by

$$H_0^{-1}(t) = \lambda^{-1}t.$$

The survival time of the Cox proportional hazards model, with a constant baseline hazard, is

$$T = \lambda^{-1}\left[-\log(U)\, exp(-\mathbf{x}'\beta)\right] = -\frac{\log(U)}{\lambda exp(\mathbf{x}'\beta)}$$

and therefore

$$T \sim Exp(\lambda e^{\mathbf{x}'\beta}).$$

In this case, the hazard function, cumulative hazard function and the survival function can be expressed in terms of the exponential distributions with the scale parameter, λ. That is,

$$h_0(t) = \lambda,$$

$$H_0(h) = \lambda t$$

and

$$S_0(t) = exp(-\lambda t).$$

The equation, $T \sim Exp(\lambda e^{\mathbf{x}'\beta})$, is used to simulate the survival time for both survival curves $S^w(t)$ and $S^c(t)$. The censoring time, $t_c \sim Exp(c)$, is assumed to be exponential with scale parameter equal to a constant rate of censoring, c, as was done in the article by Bender et al. (2005). The baseline hazard function, $h_0(t) = \lambda$, is constant and occurs when the covariate values are all equal to zero.

The formula given for the expected shortfall percentage in Section 3.3 is

$$L_{i,d}|W = \{\sigma D \times \Phi_z(D) - \sigma(-\phi_z(D))\} \times E(V_{i,d})$$

where $\Phi_z(D)$ and $\phi_z(D)$ represent the cumulative distribution function and probability density function of the standard normal distribution (Leow & Mues, 2012, p.191). The value for D will be simulated from a standard normal distribution, the valuation of the asset, $V_{i,t}$, will be simulated from a gamma distribution, and the value for σ will be taken as a constant.

The parameters in the simulation study are selected to give similar survival curves to that of a retail bank's vehicle and asset portfolio and home loans portfolio. These simulated datasets will be referred to as the simulated vehicle and asset portfolio and the simulated home loans portfolio. The parameters are varied and various survival curves are calculated. The parameter estimates used for the simulations are given in Table 2.

	Home loans		Vehicle and asset finance	
	$P_{i,t}(W_T)$	$P_{i,t}(C_T)$	$P_{i,t}(W_T)$	$P_{i,t}(C_T)$
λ	0,01	0,026	0,041	0,042
c	0,05	0,04	0,046	0,043

Table 2: *Simulation study parameters.*

The mean squared error (MSE) values between the simulated survival curves and retail portfolio survival curves are calculated. The parameter estimates of the simulated survival curve that gives the minimum MSE are used for each of the portfolios.

The loss given default for the simulated and retail vehicle and asset portfolio as well as the simulated- and retail home loans portfolio is displayed in Figure 9. The simulated- and retail LGD values for both portfolios are similar.

The distribution of the simulated and retail LGD values are displayed in Figure 10 and Figure 11. The distribution of the simulated and retail LGD compare well for the vehicle and asset finance as well as the home loans portfolio.

The probability component of the indirect LGD is predicted by making use of survival analysis (Section 3.1) and the loss severity component is predicted by the haircut model (Section 3.3). For comparative purposes, the probability component of the indirect LGD is predicted by replacing survival analysis with logistic regression. This comparison is applied to retail bank data (Section 4.1) and simulated data (Section 4.2). The MSE, bias and variance is calculated and shown in the following section.

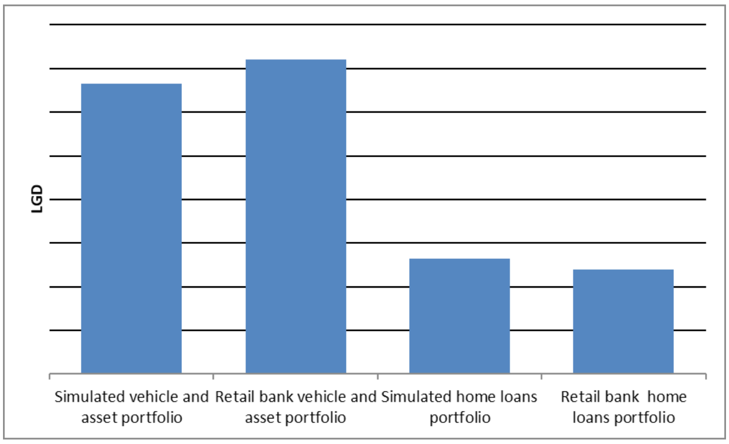


Figure 9: *Simulated loss given default.*

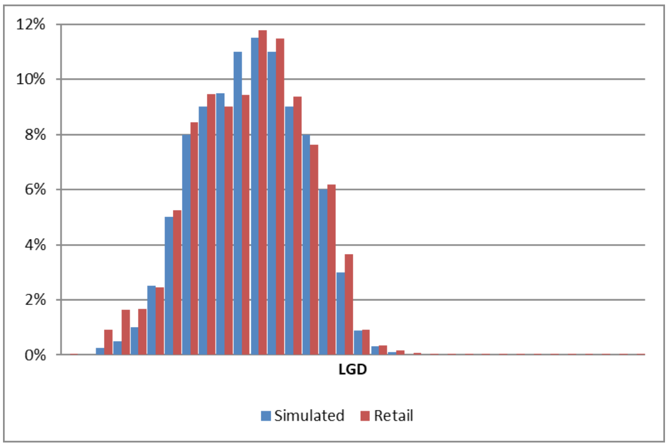


Figure 10: *Vehicle and asset finance LGD distribution.*

5 Results

This section covers the results for both the retail and the simulated datasets. The method described by Leow & Mues (2012) to predict the probability components, $P_{i,t}\left(W_{\tau}\right)$, $P_{i,t}\left(C_{\tau}\right)$ and $P_{i,t}\left(I_{\tau}\right)$, makes use of logistic regression. In the modelling methodology section, it is described how survival analysis can be used to predict these components. These probability components are combined with the haircut component to calculate LGD indirectly.

5.1 Retail Data Results

Section 5.1.1 contains accuracy graphs for each of the components. This is followed by the results for the overall LGD in Section 5.1.2.

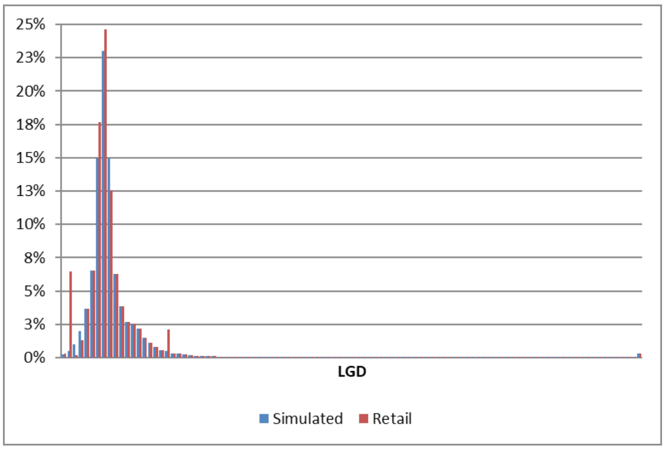


Figure 11: *Home loans LGD distribution.*

5.1.1 LGD Model Components

Accuracy graphs are displayed for the probability of cure, probability of write off and haircut models. Accounts are sorted from smallest to largest expected values and grouped into deciles that contain the same number of accounts. The actual versus expected values are given by deciles.

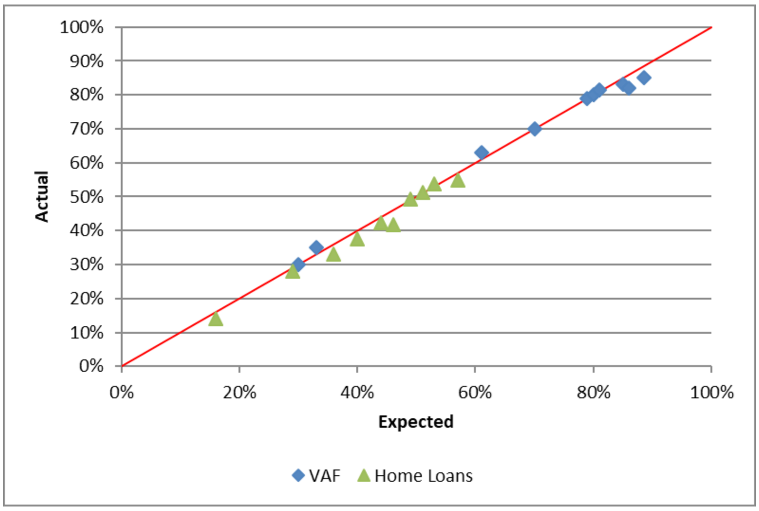


Figure 12: *Haircut model accuracy on retail data.*

One can conclude from Figure 12, Figure 13 and Figure 14 that the individual LGD model components are accurate. The accuracy graph shows that the models are accurate given that the points closely resemble a line with a 45-degree angle. The 45-degree line represents the points where the actual values are equal to the expected values. The overall LGD will be the focus in Section 5.1.2.

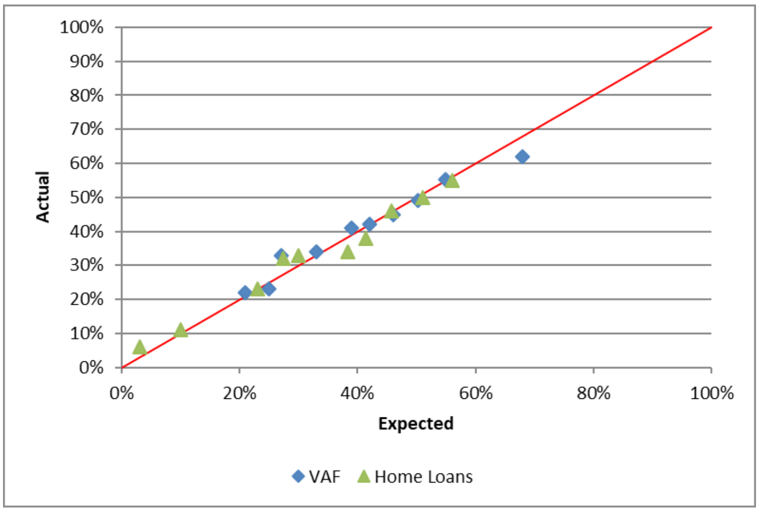


Figure 13: *Probability of cure model accuracy on retail data.*

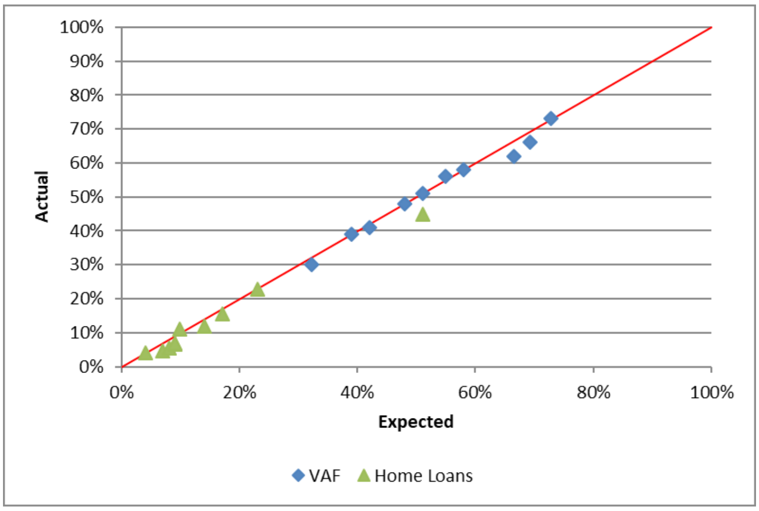


Figure 14: *Probability of write-off model accuracy on retail data.*

5.1.2 Overall LGD

Logistic regression is used to model the probability components. These components are combined with the loss severity components to estimate LGD. Logistic regression is replaced with survival analysis to estimate the probability components, but the same haircut model is used and LGD is estimated.

These LGD values are used to calculate the MSE, bias and variance on a vehicle and asset portfolio and a home loans portfolio. These values are graphically represented in Figure 15. The corresponding values are given in Table 3 and Table 4. The MSE for the survival analysis approach is the lowest in both cases and it is therefore deemed the more appropriate technique.

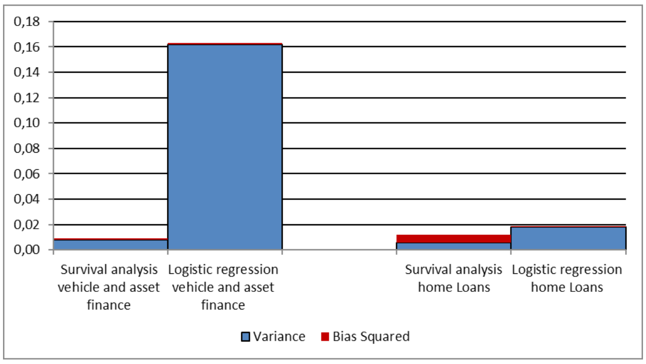


Figure 15: *Retail data MSE, variance and bias.*

The corresponding values are given in Table 3 and Table 4.

Method	Variance	MSE	Bias
Survival	0.008	0.009	-0.037
Logistic	0.162	0.163	-0.028

Table 3: *Retail data vehicle and asset portfolio.*

Method	Variance	MSE	Bias
Survival	0.006	0.012	0.077
Logistic	0.018	0.019	-0.031

Table 4: *Retail data home loans portfolio.*

In Figure 16, the expected LGD values and actual LGD values by decile are displayed for the home loans and vehicle and asset finance portfolios. The accuracy of these two models are clear from Figure 16.

The same methodology is applied to the simulated data results as used for the retail data.

5.2 Simulated Data Results

The results for the probability of cure, probability of loss and haircut models are given in Section 5.2.1 This is followed by the results for the overall LGD in Section 5.2.2.

5.2.1 LGD Model Components

Figure 17, Figure 18 and Figure 19 show that the haircut model, probability of cure model and the probability of write-off model are all accurate.

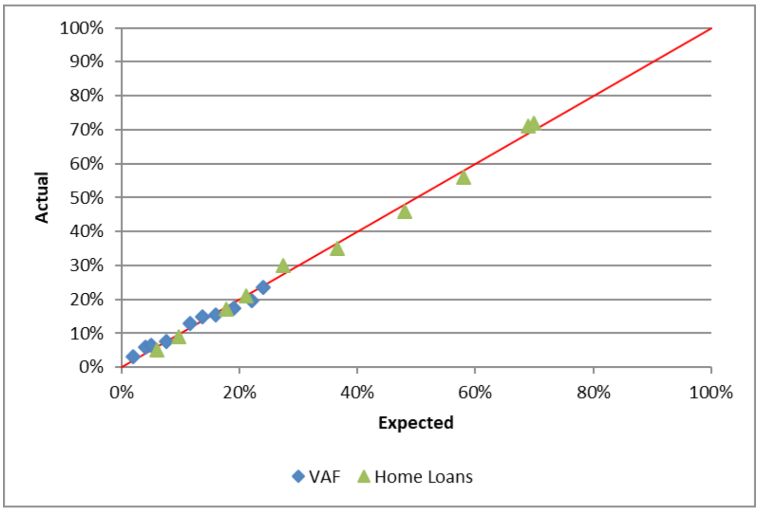


Figure 16: *LGD accuracy on retail data.*

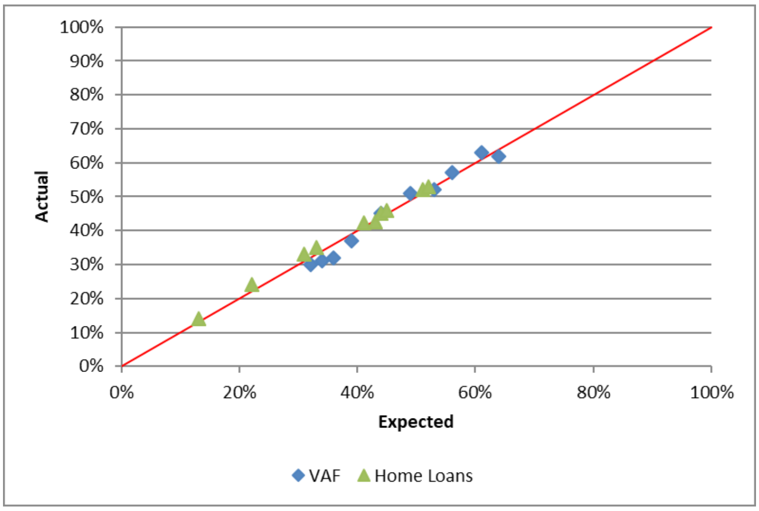


Figure 17: *Haircut model accuracy on simulated data.*

5.2.2 Overall LGD

The indirect approach is used to estimate LGD. A survival analysis approach and logistic regression approach are used to model the probability component of the LGD model. These two approaches are applied to 100,000 different simulated datasets and the MSE, bias and variance are calculated on the LGD of each set. The MSE, bias and variance are graphed in Figure 20. The corresponding values are given in Table 5 and Table 6.

The MSE for the survival analysis approach is the lowest in both cases and is therefore concluded to be the more appropriate technique. An accuracy graph, displaying the actual LGD values versus expected LGD values by decile, is displayed in Figure 21 for the

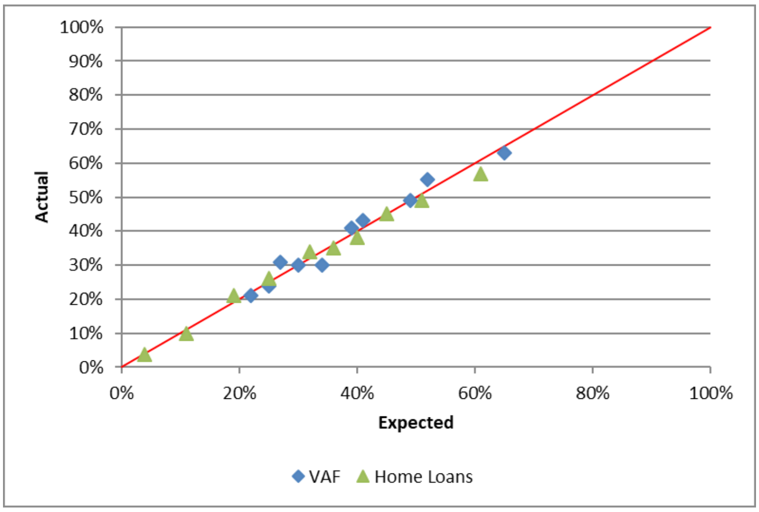


Figure 18: *Probability of cure model accuracy on simulated data.*

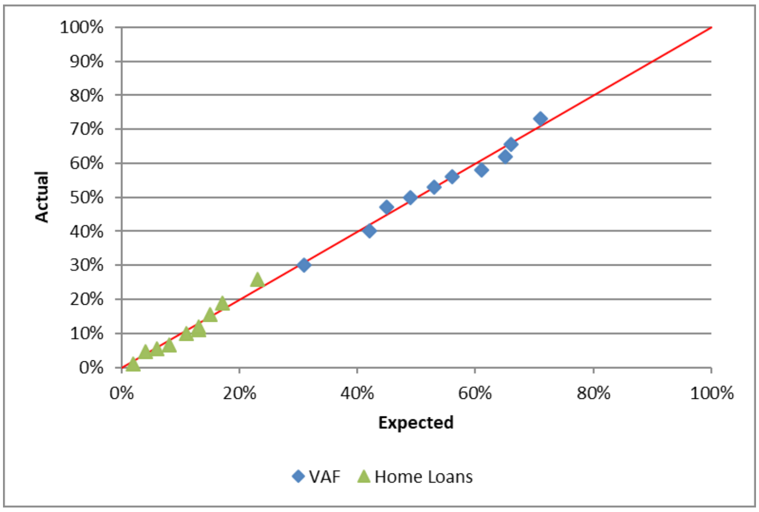


Figure 19: *Probability of write-off model accuracy on simulated data.*

Method	Variance	MSE	Bias
Survival	0.045	0.046	0.034
Logistic	0.044	0.061	-0.129

Table 5: *Simulated vehicle and asset portfolio.*

simulated home loans and vehicle and asset finance data. The accuracy of the home loans LGD and vehicle and asset finance LGD models are clear from Figure 21. The simulated dataset analysis shows that the suggested methodology generalizes well to a wide range of retail portfolios.

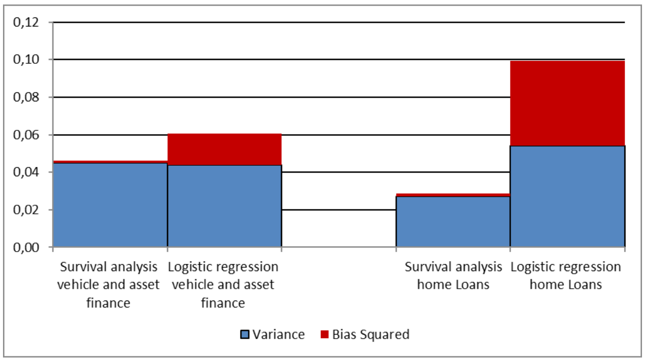


Figure 20: *Simulated MSE, variance and bias.*

Method	Variance	MSE	Bias
Survival	0.027	0.029	0.04
Logistic	0.054	0.1	0.213

Table 6: *Simulated home loans portfolio.*

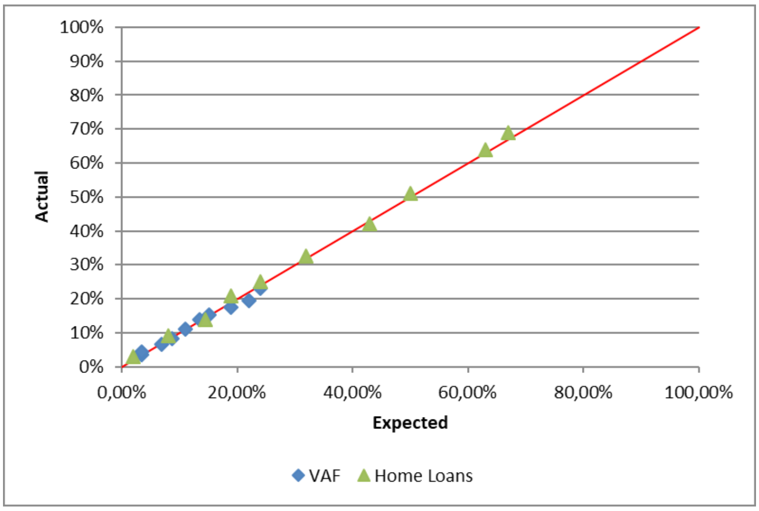


Figure 21: *LGD accuracy on simulated data.*

6 Conclusion

In a paper by Leow & Mues (2012) an indirect LGD modelling approach was described. The probability components were modelled by making use of logistic regression and the loss severity component was estimated by applying the haircut model.

This paper set out to investigate whether predictability could be enhanced if the logistic regression to model the probability component was replaced by survival analysis. Survival analysis naturally lends itself to model the probability components because it allows for censoring, a time-varying outcome window can be modelled and the survival time can be incorporated. By incorporating censoring into the models, incomplete accounts can be included in the survival analysis model. The incomplete accounts will contribute to the estimate of the survival curve up until the point where no further information is available. Having to wait for incomplete accounts to workout is now eliminated and no assumptions have to be made for these accounts. The exclusion of incomplete workouts from LGD modelling will lead to extremely inaccurate LGD values. Survival analysis allows for a varying outcome period that is advantageous since the probability to write off, probability to cure and probability to remain incomplete varies dramatically as the length of time the account is in default changes. Survival analysis makes use of the survival time and the valuable information contained in the survival time is used.

Leow & Mues (2012) modelled the probability component as a binary variable, loss or no loss. By making use of a cumulative incidence function to combine survival curves more outcomes were allowed for in our approach. The probability of cure, probability of write-off and probability of incomplete is estimated for every month that an account was in default. The expected LGD will be understated if modelling does not allow for incomplete accounts. The accuracy for each of the components used to model the LGD directly was validated and it was concluded that all these components were accurate when using our approach. The overall LGD values were also accurate. The indirect LGD was modelled on simulated datasets as well as on retail bank datasets. The MSE was compared and it was concluded that the survival analysis outperformed logistic regression.

To summarize, the contributions of this paper were that:

- The predictability of the LGD models were enhanced by making use of survival analysis instead of logistic regression in the probability components.
- Censoring was introduced to allow for accurate modelling of incomplete accounts.
- Survival time was incorporated into the probability component to enhance the information conveyed by the model.
- A time-varying outcome window instead of a fixed outcome window was incorporated to align the model to practice.
- More than two outcomes were allowed for in the target variable of the probability components to enhance accuracy.

Similar to how Miu & Ozdemir (2017) adapted Basel LGD modelling techniques to model the IFRS 9 LGD, future research could focus on extending our approach to IFRS 9 models.

References

[1] Anderson R, 2007, *The credit scoring toolkit*, Oxford university press. Great Clarendon Street, Oxford OX2 6DP.

[2] Basel committee on Banking Supervision. BCBS, 2005, *Studies on the Validation of International Rating Systems. Working paper 14*, Basel committee on Banking Supervision. Bank for International Settlements.

[3] Basel committee on Banking Supervision. BCBS, 2006, *International Convergence of Capital Measurement and Capital Standards*, Basel committee on Banking Supervision. Bank for International Settlements.

[4] Bellotti T & Crook J, 2012, *Loss given default models incorporating macroeconomic variables for credit cards*, International Journal of Forecasting 28 (2012) 171–182.

[5] Bender R, Augustin T & Blettner M, 2005, *Generating Survival Times to Simulate Cox Proportional Hazards Model*, Dept. of Epidemiology and Medical Statistics, School of Public Health University of Bielefeld, Germany, and Department of Statistics, University of Munich, German.

[6] De Jongh P, Verster T, Reynolds E, Joubert M & Raubenheimer H, 2017, *A Critical Review Of The Basel Margin Of Conservatism Requirement In A Retail Credit Context*, International Business and Economics Research Journal Fourth Quarter 2017 Volume 16, Number 4.

[7] European Banking Authority, 2017, *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures*, EBA/GL/2017/16.

[8] Engelmann B & Rauhmeier, 2011, *The Basel II Risk Parameters. Estimation, Validation, Stress Testing – with Applications to Loan Risk Management*, Springer Heidelberg Dordrecht, London/New York.

[9] Klein M & Moeschberger M, 2003, *Survival analysis techniques for censored and truncated data*, Springer-Verlag New York, Inc.

[10] Kleinbaum D & Klein M, 2012, *Survival Analysis, A Self-Learning Text, Third Edition*, Springer Science Business Media.

[11] Leow M & Mues C, 2012, *Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data*, International Journal of Forecasting, pp 183—195.

[12] Lotheram G, Brown I, Martens D, Mues C & Baesens B, 2012, *Benchmarking regression algorithms for loss given default modelling*, International Journal of Forecasting 28 (2012) 161–170.

[13] Miu P & Ozdemir B, 2017, *Adapting the Basel II advanced internal ratings-based models for International Financial Reporting Standard 9*, Journal of Credit Risk 13(2), 5383 DOI: 10.21314/JCR.2017.224.

[14] Qi M & Yang X, 2009, *Loss given default of high loan to value residential mortgages*, Journal of Banking and Finance 33 (2009) 788–799. Loss Given Default 23.

[15] Schmidt K, 2006, *Methods and Models of Loss Reserving Based on Run-Off Triangles: A Unifying Survey*, CAS 2006 Call Paper Program

[16] Somer M & Whittaker J, 2007, *Quantile regression for modelling distributions of profit and loss*, European Journal of Operational Research 183(3): 1477–1487

[17] Qi M & Zhao X, 2011, *Comparison of modelling methods for Loss Given Default*, Journal of Banking and Finance 35 (2011) 2842–2855.

[18] Tong E, Mues C & Thomas L, 2011 ,*A zero-adjusted gamma model for estimating loss given default on residential mortgage loans*, Credit Scoring and Credit Control XII, 24–26 August, Edinburgh.

[19] Tong E, Mues C & Thomas L, 2013, *A zero-adjusted gamma model for mortgage loan loss given default*, International Journal of Forecasting 29 (2013) 548–562.

[20] Witzany J, Rychnovsky M & Charamza P, 2012, *Survival Analysis in LGD Modelling*, European Financial and Accounting Journal 7 (2012) 6-27.

[21] Zhang J & Thomas L, 2012 ,*Comparisons of Linear Regression and Survival Analysis using Single and Mixture Distributions Approaches in Modelling LGD*, International Journal of Forecasting, Vol. 28, No. 1, (January-March 2012), pp. 204-215. 2012.

[22] Zhang Y, Ji L & Liu F, 2010 ,*Local housing market cycle and loss given default: Evidence from sub-prime residential mortgages*, IMF Working Paper, WP/10/167, International Monetary Fund, Washington D.C.

Chapter 3
Section 3

Errata:

Making use of Survival Analysis to indirectly model Loss Given Default.

Errata for the paper: " Making use of Survival Analysis to indirectly model Loss Given Default."

1. Page 82, Line 14 from bottom: $\lambda(t)$ change to $h(t)$.
2. The equation on Page 83, Line 8 from bottom changed to:
   $S_0\,(t) = exp\left(-\int_0^t h_0\,(s)ds\right)$
3. The equation on Page 83, Line 9 from bottom changed to:
   $$S\,(t,\mathbf{x}) = exp\left(-\int_0^t h_0\,(s)exp\left(\mathbf{x}^{'}\beta\right)ds\right) = S_0\,(t)^{exp(\mathbf{x}^{'}\beta)}\,.$$
4. Page 84, Line 6 from top:
   The likelihood function changed to:
   $$L = \prod_{i=1}^{k}\left[h_0\,(t)\,exp(\mathbf{x_i}^{'}\beta)\right]^{dN_i(t)} exp(-h_0\,(t)\,exp(\mathbf{x_i}^{'}\beta)Y_i(t)).$$
5. Page 84, Line 11 from top: The Breslow-Crowley form for the maximum likelihood estimator of the baseline hazard is changed to
   $$\hat{h}_0\,(t) = \frac{\sum\limits_{i=1}^{k} dN_i(t)}{\sum\limits_{i=1}^{k} exp(\mathbf{x_i}^{'}\beta)Y_i(t)}.$$

Chapter 4

Adapting the Default weighted survival analysis modelling approach to model the IFRS 9 LGD.

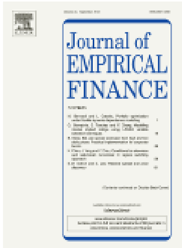
Chapter 4
Section 1

Guidelines for authors submitting an article to the Journal of Empirical Finance.

JOURNAL OF EMPIRICAL FINANCE

AUTHOR INFORMATION PACK

TABLE OF CONTENTS



ISSN: 0927-5398

DESCRIPTION

The Journal of Empirical Finance is a financial economics journal whose aim is to publish high quality articles in empirical finance. Empirical finance is interpreted broadly to include any type of empirical work in financial economics, financial econometrics, and also theoretical work with clear empirical implications, even when there is no empirical analysis. The Journal welcomes articles in all fields of finance, such as asset pricing, corporate finance, financial econometrics, banking, international finance, microstructure, behavioural finance, etc.The Editorial Team is willing to take risks on innovative research, controversial papers, and unusual approaches. We are also particularly interested in work produced by young scholars. The composition of the editorial board reflects such goals. **Editorial Policy**

We are committed to fast turnaround times. Since 2016, our goal is to make most decisions on first submissions within 10 weeks.All papers are handled by one of the main editors. For each paper, the editor chooses one of three options: The editor makes a decision on the paper without involving additional reviewers; The editor directly selects one or more ad hoc reviewers; The editor assigns the paper to an associate editor, who then selects one or more ad hoc reviewers and makes a recommendation to the editor.

In all cases, the editor is responsible for the final decision on the paper.All first submissions require payment of a submission fee. The submission fee is not refundable. In particular, the submission fee will not be refunded if the paper is "desk rejected" (i.e. the editor rejects the paper without involving additional reviewers) or if the editors are unable to secure reviewers for the paper. We do not pre-screen papers or ideas; authors have to submit their papers and pay the submission fee to receive an evaluation. Please note that, due to the exceptionally large number of high-quality submissions, the hurdle is very high: we currently reject about 85% of all submissions, of which 40% are rejected by the editors without involving further reviewers and with no detailed feedback offered. We thus recommend authors to be conservative in their submission decisions, as most submissions will lead to rejection. As a guideline for authors, here we list some of the most common reasons for desk rejections (please note that this list is not exhaustive): The paper is a better fit for Accounting, Computational, Mathematical Finance, Operations, Statistics, or Econometrics journals. The paper is a better fit for academic journals with a more practitioner orientation. The topic or the application is too narrow, being of interest to only a small group of researchers. The quality of the analysis falls short of the standards expected by the Journal. The paper is poorly written and/or formatted.

AUDIENCE

Econometricians, Financial Economists

IMPACT FACTOR

2017: 0.946 © Clarivate Analytics Journal Citation Reports 2018

ABSTRACTING AND INDEXING

RePEc
Gale Academic OneFile
Current Contents/Social & Behavioral Sciences
Current Abstracts (EBSCO)
Gale Infotrac Custom
International Bibliography of the Social Sciences
Journal of Economic Literature
Scopus
Social Sciences Citation Index

EDITORIAL BOARD

C. Julliard, London School of Economics and Political Science (LSE), London, UK
P.A.E. Koudijs, Stanford University, Stanford, California, USA
D. Li, University of South Carolina, Columbia, South Carolina, USA
D. Lou, London School of Economics and Political Science (LSE), London, England, UK
J.M. Maheu, McMaster University, Hamilton, Ontario, Canada
C. Opp, The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania, USA
L. Peng, The City University of New York, New York, USA
A. Plazzi, Università della Svizzera Italiana (USI), Lugano, Switzerland
M. Puri, Duke University, Durham, North Carolina, USA
R. Riordan, Queen's University, Ontario, Canada
D.T. Robinson, Duke University, Durham, North Carolina, USA
R. Sadka, Boston College, Chestnut Hill, Massachusetts, USA
A. Simonov, Michigan State University, East Lansing, Michigan, USA
L. Stentoft, University of Western Ontario, London, Ontario, Canada
K.S. Thorburn, Norwegian School of Economics, Bergen, Norway
R. Townsend, Dartmouth College, Hanover, New Hampshire, USA
M.A. van Dijk, Erasmus University Rotterdam, Rotterdam, Netherlands
K. Venkataraman, Southern Methodist University, Dallas, Texas, USA
J.C. Wu, University of Chicago, Chicago, Illinois, USA
Y. Xuan, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA
X. Yu, Indiana University, Bloomington, Indiana, USA
X. Zhang, Tsinghua University, Beijing, China
G. Zhou, Washington University, St. Louis, Missouri, USA

GUIDE FOR AUTHORS

The Journal of Empirical Finance will provide an international forum for empirical researchers in the intersection of the fields of econometrics and finance. The Journal welcomes high quality articles in empirical finance. Empirical finance encompasses the testing of well-established or new theories using financial data, the measurement of variables relevant in financial decision-making, the econometric methodology with finance applications. Submissions in any field of finance, corporate, international, asset pricing, market microstructure, etc. are welcome.

Submission Fees

There is a submission fee of US$ 175. Submissions will only be considered after payment of the submission fee via SubmissionStart . After you submit your manuscript, you will receive an email regarding how to send your payment. Submission fee is non-refundable and a paper may be rejected by the Editor without being sent for review, should a paper be inconsistent with the Aims and Scope of the Journal as set out on the Journal website, or not adhere to the style requirements as outlined in the Guide for Authors. The submission fees are used to support journal related activities.
There are no submission fees for submissions to a special issue.

Submission checklist

You can use this list to carry out a final check of your submission before you send it to the journal for review. Please check the relevant section in this Guide for Authors for more details.

Ensure that the following items are present:

One author has been designated as the corresponding author with contact details:
• E-mail address
• Full postal address

All necessary files have been uploaded:
Manuscript:
• Include keywords
• All figures (include relevant captions)
• All tables (including titles, description, footnotes)
• Ensure all figure and table citations in the text match the files provided
• Indicate clearly if color should be used for any figures in print
Graphical Abstracts / Highlights files (where applicable)
Supplemental files (where applicable)

Further considerations
• Manuscript has been 'spell checked' and 'grammar checked'
• All references mentioned in the Reference List are cited in the text, and vice versa
• Permission has been obtained for use of copyrighted material from other sources (including the Internet)
• A competing interests statement is provided, even if the authors have no competing interests to declare
• Journal policies detailed in this guide have been reviewed
• Referee suggestions and contact details provided, based on journal requirements

For further information, visit our Support Center.

BEFORE YOU BEGIN

Ethics in publishing

Please see our information pages on Ethics in publishing and Ethical guidelines for journal publication.

Studies in humans and animals

If the work involves the use of human subjects, the author should ensure that the work described has been carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans. The manuscript should be in line with the Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals and aim for the inclusion of representative human populations (sex, age and ethnicity) as per those recommendations. The terms sex and gender should be used correctly.

Authors should include a statement in the manuscript that informed consent was obtained for experimentation with human subjects. The privacy rights of human subjects must always be observed.

All animal experiments should comply with the ARRIVE guidelines and should be carried out in accordance with the U.K. Animals (Scientific Procedures) Act, 1986 and associated guidelines, EU Directive 2010/63/EU for animal experiments, or the National Institutes of Health guide for the care and use of Laboratory animals (NIH Publications No. 8023, revised 1978) and the authors should clearly indicate in the manuscript that such guidelines have been followed. The sex of animals must be indicated, and where appropriate, the influence (or association) of sex on the results of the study.

Declaration of interest

All authors must disclose any financial and personal relationships with other people or organizations that could inappropriately influence (bias) their work. Examples of potential competing interests include employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications/registrations, and grants or other funding. Authors must disclose any interests in two places: 1. A summary declaration of interest statement in the title page file (if double-blind) or the manuscript file (if single-blind). If there are no interests to declare then please state this: 'Declarations of interest: none'. This summary statement will be ultimately published if the article is accepted. 2. Detailed disclosures as part of a separate Declaration of Interest form, which forms part of the journal's official records. It is important for potential interests to be declared in both places and that the information matches. More information.

Submission declaration and verification

Submission of an article implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see 'Multiple, redundant or concurrent publication' section of our ethics policy for more information), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. Double submissions are not accepted and those manuscripts which will be identified as such will be immediately withdrawn; accordingly, authors are kindly required to contact the editors should they intend to withdraw their paper to submit it elsewhere. To verify originality, your article may be checked by the originality detection service CrossCheck.

Use of inclusive language

Inclusive language acknowledges diversity, conveys respect to all people, is sensitive to differences, and promotes equal opportunities. Articles should make no assumptions about the beliefs or commitments of any reader, should contain nothing which might imply that one individual is superior to another on the grounds of race, sex, culture or any other characteristic, and should use inclusive language throughout. Authors should ensure that writing is free from bias, for instance by using 'he or she', 'his/her' instead of 'he' or 'his', and by making use of job titles that are free of stereotyping (e.g. 'chairperson' instead of 'chairman' and 'flight attendant' instead of 'stewardess').

Changes to authorship

Authors are expected to consider carefully the list and order of authors **before** submitting their manuscript and provide the definitive list of authors at the time of the original submission. Any addition, deletion or rearrangement of author names in the authorship list should be made only **before** the manuscript has been accepted and only if approved by the journal Editor. To request such a change, the Editor must receive the following from the **corresponding author**: (a) the reason for the change in author list and (b) written confirmation (e-mail, letter) from all authors that they agree with the addition, removal or rearrangement. In the case of addition or removal of authors, this includes confirmation from the author being added or removed.
Only in exceptional circumstances will the Editor consider the addition, deletion or rearrangement of authors **after** the manuscript has been accepted. While the Editor considers the request, publication of the manuscript will be suspended. If the manuscript has already been published in an online issue, any requests approved by the Editor will result in a corrigendum.

Article transfer service
This journal is part of our Article Transfer Service. This means that if the Editor feels your article is more suitable in one of our other participating journals, then you may be asked to consider transferring the article to one of those. If you agree, your article will be transferred automatically on your behalf with no need to reformat. Please note that your article will be reviewed again by the new journal. More information.

Copyright

Upon acceptance of an article, authors will be asked to complete a 'Journal Publishing Agreement' (see more information on this). An e-mail will be sent to the corresponding author confirming receipt of the manuscript together with a 'Journal Publishing Agreement' form or a link to the online version of this agreement.

Subscribers may reproduce tables of contents or prepare lists of articles including abstracts for internal circulation within their institutions. Permission of the Publisher is required for resale or distribution outside the institution and for all other derivative works, including compilations and translations. If excerpts from other copyrighted works are included, the author(s) must obtain written permission from the copyright owners and credit the source(s) in the article. Elsevier has preprinted forms for use by authors in these cases.

For gold open access articles: Upon acceptance of an article, authors will be asked to complete an 'Exclusive License Agreement' (more information). Permitted third party reuse of gold open access articles is determined by the author's choice of user license.

Author rights
As an author you (or your employer or institution) have certain rights to reuse your work. More information.

Elsevier supports responsible sharing
Find out how you can share your research published in Elsevier journals.

Role of the funding source

You are requested to identify who provided financial support for the conduct of the research and/or preparation of the article and to briefly describe the role of the sponsor(s), if any, in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. If the funding source(s) had no such involvement then this should be stated.

Funding body agreements and policies
Elsevier has established a number of agreements with funding bodies which allow authors to comply with their funder's open access policies. Some funding bodies will reimburse the author for the gold open access publication fee. Details of existing agreements are available online.

Open access

This journal offers authors a choice in publishing their research:

Subscription
• Articles are made available to subscribers as well as developing countries and patient groups through our universal access programs.
• No open access publication fee payable by authors.
• The Author is entitled to post the accepted manuscript in their institution's repository and make this public after an embargo period (known as green Open Access). The published journal article cannot be shared publicly, for example on ResearchGate or Academia.edu, to ensure the sustainability of peer-reviewed research in journal publications. The embargo period for this journal can be found below.
Gold open access
• Articles are freely available to both subscribers and the wider public with permitted reuse.
• A gold open access publication fee is payable by authors or on their behalf, e.g. by their research funder or institution.

Regardless of how you choose to publish your article, the journal will apply the same peer review criteria and acceptance standards.

For gold open access articles, permitted third party (re)use is defined by the following Creative Commons user licenses:

Creative Commons Attribution (CC BY)
Lets others distribute and copy the article, create extracts, abstracts, and other revised versions, adaptations or derivative works of or from an article (such as a translation), include in a collective work (such as an anthology), text or data mine the article, even for commercial purposes, as long as they credit the author(s), do not represent the author as endorsing their adaptation of the article, and do not modify the article in such a way as to damage the author's honor or reputation.

Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)
For non-commercial purposes, lets others distribute and copy the article, and to include in a collective work (such as an anthology), as long as they credit the author(s) and provided they do not alter or modify the article.

The gold open access publication fee for this journal is **USD 1100**, excluding taxes. Learn more about Elsevier's pricing policy: https://www.elsevier.com/openaccesspricing.

Green open access
Authors can share their research in a variety of different ways and Elsevier has a number of green open access options available. We recommend authors see our green open access page for further information. Authors can also self-archive their manuscripts immediately and enable public access from their institution's repository after an embargo period. This is the version that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and in editor-author communications. Embargo period: For subscription articles, an appropriate amount of time is needed for journals to deliver value to subscribing customers before an article becomes freely available to the public. This is the embargo period and it begins from the date the article is formally published online in its final and fully citable form. Find out more.

This journal has an embargo period of 24 months.

An adapted embargo period of 12 months will apply for UK researchers who are grant recipient from the Research Council UK, Wellcome Trust, Higher Education Funding Council for England, and who wish to self-archive their accepted author manuscript. For more information, please follow this link.

Elsevier Researcher Academy
Researcher Academy is a free e-learning platform designed to support early and mid-career researchers throughout their research journey. The "Learn" environment at Researcher Academy offers several interactive modules, webinars, downloadable guides and resources to guide you through the process of writing for research and going through peer review. Feel free to use these free resources to improve your submission and navigate the publication process with ease.

Language (usage and editing services)
Please write your text in good English (American or British usage is accepted, but not a mixture of these). Authors who feel their English language manuscript may require editing to eliminate possible grammatical or spelling errors and to conform to correct scientific English may wish to use the English Language Editing service available from Elsevier's WebShop.

Submission
Our online submission system guides you stepwise through the process of entering your article details and uploading your files. The system converts your article files to a single PDF file used in the peer-review process. Editable files (e.g., Word, LaTeX) are required to typeset your article for final publication. All correspondence, including notification of the Editor's decision and requests for revision, is sent by e-mail.

Submit your article
Please submit your article via http://ees.elsevier.com/empfin

PREPARATION

Peer review
This journal operates a double blind review process. All contributions will be initially assessed by the editor for suitability for the journal. Papers deemed suitable are then typically sent to a minimum of two independent expert reviewers to assess the scientific quality of the paper. The Editor is responsible for the final decision regarding acceptance or rejection of articles. The Editor's decision is final. More information on types of peer review.

Double-blind review

This journal uses double-blind review, which means the identities of the authors are concealed from the reviewers, and vice versa. More information is available on our website. To facilitate this, please include the following separately:
Title page (with author details): This should include the title, authors' names, affiliations, acknowledgements and any Declaration of Interest statement, and a complete address for the corresponding author including an e-mail address.
Blinded manuscript (no author details): The main body of the paper (including the references, figures, tables and any acknowledgements) should not include any identifying information, such as the authors' names or affiliations.

Use of word processing software

It is important that the file be saved in the native format of the word processor used. The text should be in single-column format. Keep the layout of the text as simple as possible. Most formatting codes will be removed and replaced on processing the article. In particular, do not use the word processor's options to justify text or to hyphenate words. However, do use bold face, italics, subscripts, superscripts etc. When preparing tables, if you are using a table grid, use only one grid for each individual table and not a grid for each row. If no grid is used, use tabs, not spaces, to align columns. The electronic text should be prepared in a way very similar to that of conventional manuscripts (see also the Guide to Publishing with Elsevier). Note that source files of figures, tables and text graphics will be required whether or not you embed your figures in the text. See also the section on Electronic artwork.
To avoid unnecessary errors you are strongly advised to use the 'spell-check' and 'grammar-check' functions of your word processor.

LaTeX

You are recommended to use the Elsevier article class elsarticle.cls to prepare your manuscript and BibTeX to generate your bibliography.
Our LaTeX site has detailed submission instructions, templates and other information.

Article structure

Subdivision - numbered sections

Divide your article into clearly defined and numbered sections. Subsections should be numbered 1.1 (then 1.1.1, 1.1.2, ...), 1.2, etc. (the abstract is not included in section numbering). Use this numbering also for internal cross-referencing: do not just refer to 'the text'. Any subsection may be given a brief heading. Each heading should appear on its own separate line.

Introduction

State the objectives of the work and provide an adequate background, avoiding a detailed literature survey or a summary of the results.

Material and methods

Provide sufficient details to allow the work to be reproduced by an independent researcher. Methods that are already published should be summarized, and indicated by a reference. If quoting directly from a previously published method, use quotation marks and also cite the source. Any modifications to existing methods should also be described.

Results

Results should be clear and concise.

Discussion

This should explore the significance of the results of the work, not repeat them. A combined Results and Discussion section is often appropriate. Avoid extensive citations and discussion of published literature.

Conclusions

The main conclusions of the study may be presented in a short Conclusions section, which may stand alone or form a subsection of a Discussion or Results and Discussion section.

Appendices

If there is more than one appendix, they should be identified as A, B, etc. Formulae and equations in appendices should be given separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. Similarly for tables and figures: Table A.1; Fig. A.1, etc.

Essential title page information

• ***Title.*** Concise and informative. Titles are often used in information-retrieval systems. Avoid abbreviations and formulae where possible.
• ***Author names and affiliations.*** Please clearly indicate the given name(s) and family name(s) of each author and check that all names are accurately spelled. You can add your name between parentheses in your own script behind the English transliteration. Present the authors' affiliation addresses (where the actual work was done) below the names. Indicate all affiliations with a lower-case superscript letter immediately after the author's name and in front of the appropriate address. Provide the full postal address of each affiliation, including the country name and, if available, the e-mail address of each author.
• ***Corresponding author.*** Clearly indicate who will handle correspondence at all stages of refereeing and publication, also post-publication. This responsibility includes answering any future queries about Methodology and Materials. **Ensure that the e-mail address is given and that contact details are kept up to date by the corresponding author.**
• ***Present/permanent address.*** If an author has moved since the work described in the article was done, or was visiting at the time, a 'Present address' (or 'Permanent address') may be indicated as a footnote to that author's name. The address at which the author actually did the work must be retained as the main, affiliation address. Superscript Arabic numerals are used for such footnotes.

Abstract

A concise and factual abstract is required. The abstract should state briefly the purpose of the research, the principal results and major conclusions. An abstract is often presented separately from the article, so it must be able to stand alone. For this reason, References should be avoided, but if essential, then cite the author(s) and year(s). Also, non-standard or uncommon abbreviations should be avoided, but if essential they must be defined at their first mention in the abstract itself.

Highlights
Highlights are mandatory for this journal. They consist of a short collection of bullet points that convey the core findings of the article and should be submitted in a separate editable file in the online submission system. Please use 'Highlights' in the file name and include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point). You can view example Highlights on our information site.

Keywords

Immediately after the abstract, provide a maximum of 6 keywords, using American spelling and avoiding general and plural terms and multiple concepts (avoid, for example, 'and', 'of'). Be sparing with abbreviations: only abbreviations firmly established in the field may be eligible. These keywords will be used for indexing purposes.

Classification codes
Please provide up to 6 standard JEL codes. The available codes may be accessed at JEL.

Abbreviations
Define abbreviations that are not standard in this field in a footnote to be placed on the first page of the article. Such abbreviations that are unavoidable in the abstract must be defined at their first mention there, as well as in the footnote. Ensure consistency of abbreviations throughout the article.

Acknowledgements
Collate acknowledgements in a separate section at the end of the article before the references and do not, therefore, include them on the title page, as a footnote to the title or otherwise. List here those individuals who provided help during the research (e.g., providing language help, writing assistance or proof reading the article, etc.).

Formatting of funding sources
List funding sources in this standard way to facilitate compliance to funder's requirements:

Funding: This work was supported by the National Institutes of Health [grant numbers xxxx, yyyy]; the Bill & Melinda Gates Foundation, Seattle, WA [grant number zzzz]; and the United States Institutes of Peace [grant number aaaa].

It is not necessary to include detailed descriptions on the program or type of grants and awards. When funding is from a block grant or other resources available to a university, college, or other research institution, submit the name of the institute or organization that provided the funding.

If no funding has been provided for the research, please include the following sentence:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Math formulae
Please submit math equations as editable text and not as images. Present simple formulae in line with normal text where possible and use the solidus (/) instead of a horizontal line for small fractional terms, e.g., X/Y. In principle, variables are to be presented in italics. Powers of e are often more conveniently denoted by exp. Number consecutively any equations that have to be displayed separately from the text (if referred to explicitly in the text).

Footnotes
Footnotes should be used sparingly. Number them consecutively throughout the article. Many word processors can build footnotes into the text, and this feature may be used. Otherwise, please indicate the position of footnotes in the text and list the footnotes themselves separately at the end of the article. Do not include footnotes in the Reference list.

Artwork

Electronic artwork
General points
• Make sure you use uniform lettering and sizing of your original artwork.
• Embed the used fonts if the application provides that option.
• Aim to use the following fonts in your illustrations: Arial, Courier, Times New Roman, Symbol, or use fonts that look similar.
• Number the illustrations according to their sequence in the text.
• Use a logical naming convention for your artwork files.
• Provide captions to illustrations separately.
• Size the illustrations close to the desired dimensions of the published version.
• Submit each illustration as a separate file.
A detailed guide on electronic artwork is available.
You are urged to visit this site; some excerpts from the detailed information are given here.
Formats
If your electronic artwork is created in a Microsoft Office application (Word, PowerPoint, Excel) then please supply 'as is' in the native document format.
Regardless of the application used other than Microsoft Office, when your electronic artwork is finalized, please 'Save as' or convert the images to one of the following formats (note the resolution requirements for line drawings, halftones, and line/halftone combinations given below):
EPS (or PDF): Vector drawings, embed all used fonts.
TIFF (or JPEG): Color or grayscale photographs (halftones), keep to a minimum of 300 dpi.
TIFF (or JPEG): Bitmapped (pure black & white pixels) line drawings, keep to a minimum of 1000 dpi.
TIFF (or JPEG): Combinations bitmapped line/half-tone (color or grayscale), keep to a minimum of 500 dpi.
Please do not:
• Supply files that are optimized for screen use (e.g., GIF, BMP, PICT, WPG); these typically have a low number of pixels and limited set of colors;
• Supply files that are too low in resolution;
• Submit graphics that are disproportionately large for the content.

Color artwork
Please make sure that artwork files are in an acceptable format (TIFF (or JPEG), EPS (or PDF), or MS Office files) and with the correct resolution. If, together with your accepted article, you submit usable color figures then Elsevier will ensure, at no additional charge, that these figures will appear in color online (e.g., ScienceDirect and other sites) regardless of whether or not these illustrations are reproduced in color in the printed version. **For color reproduction in print, you will receive information regarding the costs from Elsevier after receipt of your accepted article**. Please indicate your preference for color: in print or online only. Further information on the preparation of electronic artwork.

Figure captions
Ensure that each illustration has a caption. Supply captions separately, not attached to the figure. A caption should comprise a brief title (**not** on the figure itself) and a description of the illustration. Keep text in the illustrations themselves to a minimum but explain all symbols and abbreviations used.

Tables

Please submit tables as editable text and not as images. Tables can be placed either next to the relevant text in the article, or on separate page(s) at the end. Number tables consecutively in accordance with their appearance in the text and place any table notes below the table body. Be sparing in the use of tables and ensure that the data presented in them do not duplicate results described elsewhere in the article. Please avoid using vertical rules and shading in table cells.

References

Citation in text

Please ensure that every reference cited in the text is also present in the reference list (and vice versa). Any references cited in the abstract must be given in full. Unpublished results and personal communications are not recommended in the reference list, but may be mentioned in the text. If these references are included in the reference list they should follow the standard reference style of the journal and should include a substitution of the publication date with either 'Unpublished results' or 'Personal communication'. Citation of a reference as 'in press' implies that the item has been accepted for publication.

Reference links

Increased discoverability of research and high quality peer review are ensured by online links to the sources cited. In order to allow us to create links to abstracting and indexing services, such as Scopus, CrossRef and PubMed, please ensure that data provided in the references are correct. Please note that incorrect surnames, journal/book titles, publication year and pagination may prevent link creation. When copying references, please be careful as they may already contain errors. Use of the DOI is highly encouraged.

A DOI is guaranteed never to change, so you can use it as a permanent link to any electronic article. An example of a citation using DOI for an article not yet in an issue is: VanDecar J.C., Russo R.M., James D.E., Ambeh W.B., Franke M. (2003). Aseismic continuation of the Lesser Antilles slab beneath northeastern Venezuela. Journal of Geophysical Research, https://doi.org/10.1029/2001JB000884. Please note the format of such citations should be in the same style as all other references in the paper.

Web references

As a minimum, the full URL should be given and the date when the reference was last accessed. Any further information, if known (DOI, author names, dates, reference to a source publication, etc.), should also be given. Web references can be listed separately (e.g., after the reference list) under a different heading if desired, or can be included in the reference list.

Data references

This journal encourages you to cite underlying or relevant datasets in your manuscript by citing them in your text and including a data reference in your Reference List. Data references should include the following elements: author name(s), dataset title, data repository, version (where available), year, and global persistent identifier. Add [dataset] immediately before the reference so we can properly identify it as a data reference. The [dataset] identifier will not appear in your published article.

References in a special issue

Please ensure that the words 'this issue' are added to any references in the list (and any citations in the text) to other articles in the same Special Issue.

Reference management software

Most Elsevier journals have their reference template available in many of the most popular reference management software products. These include all products that support Citation Style Language styles, such as Mendeley and Zotero, as well as EndNote. Using the word processor plug-ins from these products, authors only need to select the appropriate journal template when preparing their article, after which citations and bibliographies will be automatically formatted in the journal's style. If no template is yet available for this journal, please follow the format of the sample references and citations as shown in this Guide. If you use reference management software, please ensure that you remove all field codes before submitting the electronic manuscript. More information on how to remove field codes.

Users of Mendeley Desktop can easily install the reference style for this journal by clicking the following link:
http://open.mendeley.com/use-citation-style/journal-of-empirical-finance
When preparing your manuscript, you will then be able to select this style using the Mendeley plug-ins for Microsoft Word or LibreOffice.

Reference formatting
There are no strict requirements on reference formatting at submission. References can be in any style or format as long as the style is consistent. Where applicable, author(s) name(s), journal title/ book title, chapter title/article title, year of publication, volume number/book chapter and the article number or pagination must be present. Use of DOI is highly encouraged. The reference style used by the journal will be applied to the accepted article by Elsevier at the proof stage. Note that missing data will be highlighted at proof stage for the author to correct. If you do wish to format the references yourself they should be arranged according to the following examples:

Reference style
Text: All citations in the text should refer to:
1. *Single author:* the author's name (without initials, unless there is ambiguity) and the year of publication;
2. *Two authors:* both authors' names and the year of publication;
3. *Three or more authors:* first author's name followed by 'et al.' and the year of publication.
Citations may be made directly (or parenthetically). Groups of references can be listed either first alphabetically, then chronologically, or vice versa.
Examples: 'as demonstrated (Allan, 2000a, 2000b, 1999; Allan and Jones, 1999)…. Or, as demonstrated (Jones, 1999; Allan, 2000)… Kramer et al. (2010) have recently shown …'
List: References should be arranged first alphabetically and then further sorted chronologically if necessary. More than one reference from the same author(s) in the same year must be identified by the letters 'a', 'b', 'c', etc., placed after the year of publication.
Examples:
Reference to a journal publication:
Van der Geer, J., Hanraads, J.A.J., Lupton, R.A., 2010. The art of writing a scientific article. J. Sci. Commun. 163, 51–59. https://doi.org/10.1016/j.Sc.2010.00372.
Reference to a journal publication with an article number:
Van der Geer, J., Hanraads, J.A.J., Lupton, R.A., 2018. The art of writing a scientific article. Heliyon. 19, e00205. https://doi.org/10.1016/j.heliyon.2018.e00205.
Reference to a book:
Strunk Jr., W., White, E.B., 2000. The Elements of Style, fourth ed. Longman, New York.
Reference to a chapter in an edited book:
Mettam, G.R., Adams, L.B., 2009. How to prepare an electronic version of your article, in: Jones, B.S., Smith , R.Z. (Eds.), Introduction to the Electronic Age. E-Publishing Inc., New York, pp. 281–304.
Reference to a website:
Cancer Research UK, 1975. Cancer statistics reports for the UK. http://www.cancerresearchuk.org/ aboutcancer/statistics/cancerstatsreport/ (accessed 13 March 2003).
Reference to a dataset:
[dataset] Oguro, M., Imahiro, S., Saito, S., Nakashizuka, T., 2015. Mortality data for Japanese oak wilt disease and surrounding forest compositions. Mendeley Data, v1. https://doi.org/10.17632/ xwj98nb39r.1.

Journal abbreviations source
Journal names should be abbreviated according to the List of Title Word Abbreviations.

Video
Elsevier accepts video material and animation sequences to support and enhance your scientific research. Authors who have video or animation files that they wish to submit with their article are strongly encouraged to include links to these within the body of the article. This can be done in the same way as a figure or table by referring to the video or animation content and noting in the body text where it should be placed. All submitted files should be properly labeled so that they directly relate to the video file's content. . In order to ensure that your video or animation material is directly usable, please provide the file in one of our recommended file formats with a preferred maximum size of 150 MB per file, 1 GB in total. Video and animation files supplied will be published online in the electronic version of your article in Elsevier Web products, including ScienceDirect. Please supply 'stills' with your files: you can choose any frame from the video or animation or make a separate image. These will be used instead of standard icons and will personalize the link to your video data. For more detailed instructions please visit our video instruction pages. Note: since video and animation cannot be embedded in the print version of the journal, please provide text for both the electronic and the print version for the portions of the article that refer to this content.

Data visualization
Include interactive data visualizations in your publication and let your readers interact and engage more closely with your research. Follow the instructions here to find out about available data visualization options and how to include them with your article.

Supplementary material
Supplementary material such as applications, images and sound clips, can be published with your article to enhance it. Submitted supplementary items are published exactly as they are received (Excel or PowerPoint files will appear as such online). Please submit your material together with the article and supply a concise, descriptive caption for each supplementary file. If you wish to make changes to supplementary material during any stage of the process, please make sure to provide an updated file. Do not annotate any corrections on a previous version. Please switch off the 'Track Changes' option in Microsoft Office files as these will appear in the published version.

Research data
This journal encourages and enables you to share data that supports your research publication where appropriate, and enables you to interlink the data with your published articles. Research data refers to the results of observations or experimentation that validate research findings. To facilitate reproducibility and data reuse, this journal also encourages you to share your software, code, models, algorithms, protocols, methods and other useful materials related to the project.

Below are a number of ways in which you can associate data with your article or make a statement about the availability of your data when submitting your manuscript. If you are sharing data in one of these ways, you are encouraged to cite the data in your manuscript and reference list. Please refer to the "References" section for more information about data citation. For more information on depositing, sharing and using research data and other relevant research materials, visit the research data page.

Data linking
If you have made your research data available in a data repository, you can link your article directly to the dataset. Elsevier collaborates with a number of repositories to link articles on ScienceDirect with relevant repositories, giving readers access to underlying data that gives them a better understanding of the research described.

There are different ways to link your datasets to your article. When available, you can directly link your dataset to your article by providing the relevant information in the submission system. For more information, visit the database linking page.

For supported data repositories a repository banner will automatically appear next to your published article on ScienceDirect.

In addition, you can link to relevant data or entities through identifiers within the text of your manuscript, using the following format: Database: xxxx (e.g., TAIR: AT1G01020; CCDC: 734053; PDB: 1XFN).

Mendeley Data
This journal supports Mendeley Data, enabling you to deposit any research data (including raw and processed data, video, code, software, algorithms, protocols, and methods) associated with your manuscript in a free-to-use, open access repository. During the submission process, after uploading your manuscript, you will have the opportunity to upload your relevant datasets directly to *Mendeley Data*. The datasets will be listed and directly accessible to readers next to your published article online.

For more information, visit the Mendeley Data for journals page.

Data statement
To foster transparency, we encourage you to state the availability of your data in your submission. This may be a requirement of your funding body or institution. If your data is unavailable to access or unsuitable to post, you will have the opportunity to indicate why during the submission process, for example by stating that the research data is confidential. The statement will appear with your published article on ScienceDirect. For more information, visit the Data Statement page.

AFTER ACCEPTANCE

Online proof correction
Corresponding authors will receive an e-mail with a link to our online proofing system, allowing annotation and correction of proofs online. The environment is similar to MS Word: in addition to editing text, you can also comment on figures/tables and answer questions from the Copy Editor. Web-based proofing provides a faster and less error-prone process by allowing you to directly type your corrections, eliminating the potential introduction of errors.
If preferred, you can still choose to annotate and upload your edits on the PDF version. All instructions for proofing will be given in the e-mail we send to authors, including alternative methods to the online version and PDF.
We will do everything possible to get your article published quickly and accurately. Please use this proof only for checking the typesetting, editing, completeness and correctness of the text, tables and figures. Significant changes to the article as accepted for publication will only be considered at this stage with permission from the Editor. It is important to ensure that all corrections are sent back to us in one communication. Please check carefully before replying, as inclusion of any subsequent corrections cannot be guaranteed. Proofreading is solely your responsibility.

Offprints
The corresponding author will, at no cost, receive a customized Share Link providing 50 days free access to the final published version of the article on ScienceDirect. The Share Link can be used for sharing the article via any communication channel, including email and social media. For an extra charge, paper offprints can be ordered via the offprint order form which is sent once the article is accepted for publication. Both corresponding and co-authors may order offprints at any time via Elsevier's Webshop. Corresponding authors who have published their article gold open access do not receive a Share Link as their final published version of the article is available open access on ScienceDirect and can be shared through the article DOI link.

AUTHOR INQUIRIES
Visit the Elsevier Support Center to find the answers you need. Here you will find everything from Frequently Asked Questions to ways to get in touch.
You can also check the status of your submitted article or find out when your accepted article will be published.

Chapter 4
Section 2

Article Title:

Adapting the Default weighted survival analysis modelling approach to model the IFRS 9 LGD.

Article Authors:

M. Joubert, T. Verster and H. Raubenheimer.

The article was submitted for publication to the Journal of Empirical Finance (2018).

Contents of Chapter 4 Section 2

List of Figures

List of Tables

ADAPTING THE DEFAULT WEIGHTED SURVIVAL ANALYSIS MODELLING APPROACH TO MODEL THE IFRS 9 LGD

blind referee copy [1]

Key words: Loss Given Default, Survival Analysis, IFRS 9.

Abstract:

Survival analysis is one of several methods used to predict loss given default (LGD) for Basel regulatory purposes. When using survival analysis to model LGD, one proposed methodology is the default weighted survival analysis (DWSA) method. This paper is aimed at adapting the DWSA method used to model the Basel LGD to estimate the LGD for International Financial Reporting Standard (IFRS) 9 impairment requirements. The DWSA methodology allows for over recoveries, default weighting and negative cashflows. This needs to be adapted for IFRS 9. This IFRS 9 LGD is used in the calculation of the expected credit losses (ECL) as per the IFRS 9 standard. The IFRS 9 LGD methodology that is described in this paper makes use of survival analysis to estimate the LGD. The Cox proportional hazards model allows for a baseline survival curve to be adjusted to produce survival curves for different segments of the portfolio. The forward-looking LGD values are adjusted for different macro-economic scenarios and an ECL is calculated for each scenario. These ECL values are probability weighted to produce a single ECL number. This paper illustrates the IFRS 9 LGD as well as the ECL on a real dataset from a retail portfolio of a South African bank.

1. Introduction

Joubert, Verster and Raubenheimer (2018) propose a default weighted survival analysis (DWSA) to directly model loss given default (LGD). This DWSA methodology was motivated by Basel considerations. Alternatives to the DWSA methodology is to model Basel LGD by making use of run-off triangles (Braun, 2004, pp. 401), beta regression (Brown, 2014, pp. 65 – 66), ordinary least squared approach (Witzany, Rychnovsky and Charamza, 2012, pp. 12), fractional response regression (Bastos, 2010, pp. 2512) and inverse beta model (Brown, 2014, pp. 64).

[1] Corresponding author.

In this paper, the DWSA methodology is adapted to model the International Financial Reporting Standard (IFRS) 9 LGD. The DWSA methodology models the Basel LGD. First we will discuss some background in Section 1.1 and then specifically refer to IFRS 9 and the specific concepts that need to be addressed to be able to adapt the DWSA methodology to be used for IFRS 9. Many adaptations will be required. For example, the main segmentation used in the DWSA model is the default date and months since default. This needs to be changed to be suitable for IFRS 9. Note that the IFRS 9 LGD models are primarily point in time models and are calibrated to recent information (Chawla, Forest and Aguais, 2016). The lifetime of the account forms a primary part of the expected credit loss calculation, and the application date and months on book form the main segmentation for the IFRS 9 LGD model (for Basel the main segmentation is month since default and default date). LGD is modelled by month on book and a separate survival curve is created for every month on book. The DWSA methodology is used to create each of these survival curves and the segmentation for IFRS 9 changes from the default date and months since default to the application date and months on book. To summarize, this paper provides a description of IFRS 9 concepts and calculations and introduce a new IFRS 9 LGD methodology. The new IFRS 9 LGD methodology is described and fitted to a retail banking dataset. The forward-looking IFRS 9 LGD is adjusted for macro-economic scenarios by applying the error correction model (ECM). The adaptations can be summarized under the following headings: segmentation, reference period, modelling population, LGD calculations, incorporating over recoveries into the IFRS 9 LGD, and macro-economic adjustment. SAS software is used for the statistical analysis, model development and data manipulation relating to this article.

1.1. Background

The International Accounting Standard Board published the new and complete IFRS 9 standard in the form of the document titled IFRS 9 Financial Instruments (IFRS, 2014). This document replaces most of the International Accounting Standard (IAS) 39. It contains impairment requirements that allows for earlier recognition of credit losses. The financial statements of banks must reflect the IFRS 9 accounting standards from 1 January 2018 (European Banking Authority (EBA), 2016, pp. 4). Banks are expecting IFRS 9 to have a significant impact on systems and processes (Beerbaum, 2015). The IAS 39 makes use of provisions on incurred losses. Learning gained from the financial crisis showed that expected losses, instead of incurred losses, should be used to calculate the provisioning for banks (Global Public Policy Committee (GPPC), 2016, pp. 21). Under IFRS 9, the expected credit losses (ECL) should be equal to an amount equivalent to the lifetime ECL, if the credit risk has risen significantly. When the converse is true, a financial entity may allow for credit losses equal to 12-month ECL. The ECL model is a forward-looking model and should result in the early detection of credit losses. This will contribute to financial stability (IFRS, 2014, pp. 26).

The ECL for account i, that is currently at month on book m, is calculated as:

$$ECL_{i,m} = \sum\nolimits_{h=0}^{H} \frac{PD_{i,m,m+h} LGD_{i,m+h} EAD_{i,m+h}}{(1+e)^h}.$$

The marginal $PD_{i,m,m+h}$ is the probability of account i defaulting at month on book $m+h$, given that the account remained performing until month on book m. $LGD_{i,m+h}$ is the loss given that account i defaults at month on book $m+h$, and $EAD_{i,m+h}$ is the exposure of account i that defaulted at month on book $m+h$. The value e is the monthly effective interest rate. LGD is always assessed over the life of the lending exposure (Basel Committee on Banking Supervision (BCBS), 2015, pp. 25). The length of the future time horizon, H, for the forward-looking information used in the estimation of ECL will vary between 12 months or the remaining lifetime, depending on the stage that an account is in. Different modelling approaches are followed for accounts in different **stages**. An account can either fall into stage 1, stage 2 or stage 3. Stage 1 accounts are performing accounts. Stage 2 accounts have significant deterioration in credit risk, but are not in default. Defaulted accounts are in stage 3 (Aptivaa, 2016a). The value of H for stage 1 and stage 2 will be equal to 12 and $max(remaining term, 12)$, respectively.

1.2. IFRS 9 concepts

The following concepts form the basis of the discussions in the IFRS 9 standard and will be used in this paper: significant deterioration, default definition, staging, lifetime, forward-looking, macro-economic factors and time value of money.

A thirty days past due rule is suggested as a "backstop" when determining if an account has **significantly deteriorated** since origination (IFRS, 2014, pp. 27). The changes in an account's behavioural score or bureau score, when measured from origination, can be used to determine significant deterioration. Accounts that are thirty days past due will be flagged as significantly deteriorated and the scores are used to select significantly deteriorated accounts from the population that are not thirty days past due.

The default point will occur after the point of significant deterioration. A 90 day past due rule is suggested as a backstop for default and the **default definition** should be in line with default definitions already used for risk management purposes (IFRS, 2014, pp. 120). The Basel default definition flags 90 days past due as default and is already deeply embedded in risk management frameworks. The IFRS 9 default definition is aligned with the Basel default definition.

The **lifetime** of an account is taken as the maximum contractual period over which the accounts are exposed to credit risk (IFRS, 2014, pp. 120). For fixed term products, the term of the account can be taken as the lifetime of the accounts, but it is important to recognize all recoveries on the accounts after the write-off point. An analysis should be conducted to measure the amount recovered after the end of the contractual term of the accounts to determine when the lifetime of the account ends. All available information is taken into account when measuring the lifetime of an account that does not have a fixed term. This approach will only work in environments where sufficient history on accounts exist. The history built up and used for Basel model development purposes should be utilized.

Forward-looking information is used when modelling ECL (Aptivaa, 2016b). The expected loss or the components to determine expected credit losses are forecasted for future time periods, and the sum over the current and future period is taken to predict expected losses (Miu and Ozdemir, 2017).

Macro-economic factors are modelled onto the time series and different economic scenarios are expressed in terms of the factors included in this model, and a probability is assigned to each of these scenarios (Black, Levine and Licari, 2015). The scenarios are applied to accounts through the macro-economic factors and a probability weighted average expected loss per account is calculated (IFRS, 2014, pp. 121).

The **time value of money** is considered when calculating the expected credit loss. The cashflows on accounts are discounted to the reporting date by applying the current monthly effective interest rate (e) (IFRS, 2014, pp. 122).

The paper is structured in the following way. The modelling approach is described in Section 2. It starts off by describing the DWSA modelling approach and how this approach is used to model the Basel LGD. The adaptation made to the DWSA approach to model the IFRS 9 LGD is described. The data is described in Section 3 and the results given in Section 4. Section 5 concludes.

2. Modelling methodology

The DWSA model (Joubert et al., 2018) estimates the Basel LGD. This approach is adapted to estimate the IFRS 9 LGD. A description of the DWSA model is given in Section 2.1 and the adaptations made to the DWSA model to comply with IFRS 9 is given in Section 2.2.

2.1. Default weighted survival analysis

The details of the DWSA methodology can be found in Joubert et al. (2018). This section is only a summary of the DWSA methodology. Mathematically the LGD at time of default may be expressed as:

$$LGD_i = \frac{EAD_i - \sum_{t=1}^{T_w} DCF_{i,t}}{EAD_i},$$

where $DCF_{i,t} = \frac{CF_{i,t}}{(1+e)^t}$ is the discounted future cashflows for account i at time t and EAD_i the EAD for account i. The recovery time, t, for a defaulted account, i, is measured in months and takes only values $1, 2, ..., T_w$. Cashflows, $CF_{i,t}$ are calculated as the difference between account balances now versus account balances in the previous month, adding back the interest and the fees, subtracting the amount written off. The post write-off recoveries represent recovery or additional expense amounts post the write-off date, which are added to the cashflows (Witzany et al., 2012, pp. 8). All accounts with recovery information up until time T_w are deemed to have complete recovery information. The time when the recovery process ends for account i will be denoted by $t_{i,end}$. The recovery process is completed if $t_{i,end} < T_w$ for example if an account closes before T_w. Alternatively the recovery process is incomplete if $T_w \leq t_{i,end}$.

A survival curve, $S(t,i) = 1 - P(T < t)$ is defined as the (unrecovered) proportion of EAD_i that remains in default up to a specific recovery time t, where $t \in \{1, \ldots, T_w\}$ for account i. Thus:

$$S(t,i) = \frac{EAD_i - \sum_{s=1}^{t} DCF_{i,s}}{EAD_i}.$$

The Kaplan-Meier estimate, $\hat{S}(t,i)$, is the empirical value of the survival curve calculated from the data and is equal to:

$$\hat{S}(t,i) = \frac{E\hat{A}D_i - \sum_{s=1}^{t} D\hat{C}F_{i,s}}{E\hat{A}D_i},$$

where $\hat{EAD}_i$ is the exposure at the default point for account i and $D\hat{C}F_{i,s}$ is the value of the cashflow for account i at point s. The survival curve for the population is then calculate as:

$$\hat{S}_0(t) = \frac{E\hat{A}D - \sum_{s=1}^{t} D\hat{C}F_s}{E\hat{A}D},$$

where $E\hat{A}D = \sum_i E\hat{A}D_i$ and $D\hat{C}F_s = \sum_i D\hat{C}F_{i,s}$.

The data needs to be constructed in a specific way when applying the DWSA methodology to incorporate censoring, incomplete records, unrecovered amounts, etc. For more information on implementation and dataset construction see (Joubert et al., 2018).

The general form of the Cox model can be written as:

$$S(t, \boldsymbol{x}_i) = S_0(t)^{exp(\boldsymbol{x}_{\mathrm{i}}{}'\beta)},$$

where $\mathbf{x_i}$ are the covariate values for account i. The weighted survival curve at time t in default contains a component $S_0(t)$ that is known as the baseline survival curve. The baseline survival curve is the Kaplan–Meier estimate of the portfolio. If an account falls outside of the base, the baseline, $S_0(t)$, is shifted by the exponent of $exp(\boldsymbol{x}_{\mathrm{i}}{}'\beta)$. The loss given default for account i at default is calculated as:

$$LGD_i = \frac{S(T_w, \boldsymbol{x}_i)}{S(0, \boldsymbol{x}_i)}.$$

The DWSA modelling methodology models the Basel LGD. This methodology is adapted to model the IFRS 9 LGD. These adaptations are described in Section 2.2.

2.2. Adaptations

The adaptations to the DWSA approach are described under the headings segmentation, reference period, LGD calculation and macro-economic model.

2.2.1. Segmentation

The IFRS accord stipulates that forward-looking information must be used when estimating the IFRS 9 LGD. LGD_m, the loss given that the account defaults at month on book m, is calculated from historical data. The segmentation for the DWSA approach is adapted from using the default date and months since default, to using the month on book, m, and the application date.

2.2.2. Reference period

The IFRS 9 LGD is calibrated to recent information. The reference period for the IFRS 9 LGD is illustrated in Figure 1 and falls between the two blue dotted lines where the information for the months January 2016 to December 2017 is used. The example in Figure 1 illustrates how the population for $m = 4$ is selected. All accounts that default at $m = 4$ (see for example account A and C in Figure 1) are considered. The exposure at entry into the reference period and the cashflows within the reference period, for accounts defaulting at $m = 4$ are considered to calculate LGD_4.

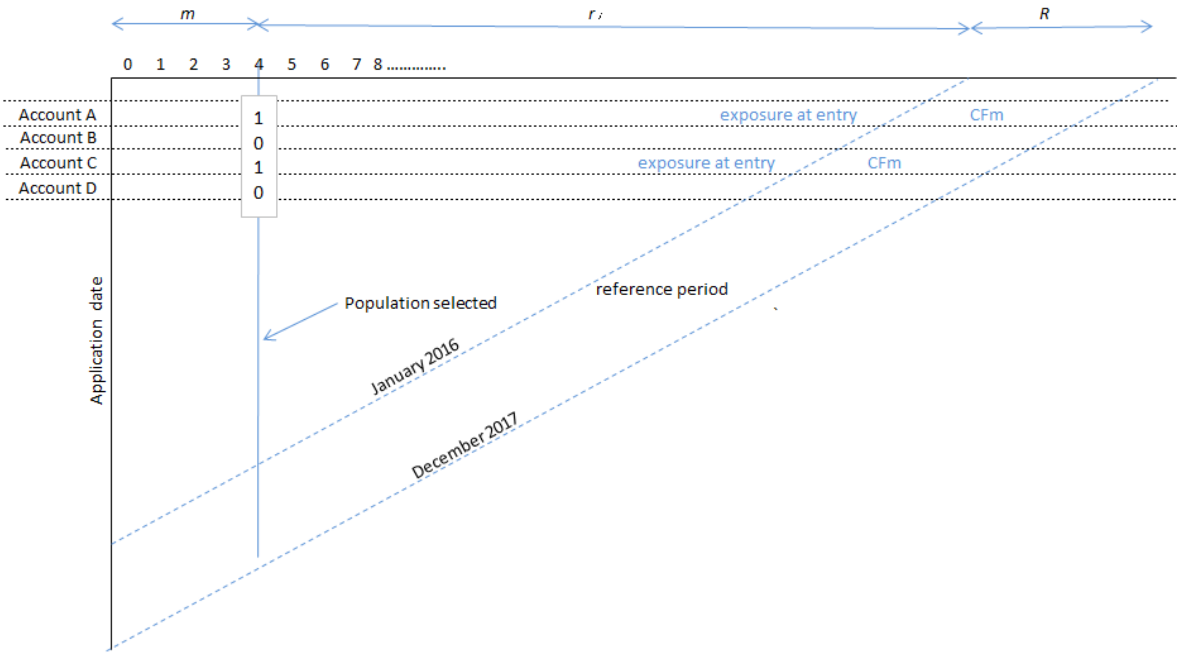


Figure 1: IFRS 9 reference period

The exposure at entry into the reference period is used to model the IFRS LGD; this replaces the exposure at default used for modelling the Basel LGD. The cashflows used for the Basel LGD span the workout period. The cashflows for the IFRS LGD span the recent reference period as displayed in Figure 1.

2.2.3. LGD calculation

The Basel loss given default under the DWSA approach for account i at default may be calculated as:

$$LGD_i = \frac{S(T_w, \boldsymbol{x}_i)}{S(0, \boldsymbol{x}_i)}.$$

The IFRS 9 LGD for account i that defaults at a specific month on book m is:

$$LGD_{i,m} = \frac{DE_{i,m} - \sum_{n=m+r_i+1}^{m+r_i+R} DCF_{i,n}}{DE_{i,m}},$$

where $DE_{i,m} = EAD_{i,m} - \sum_{n=m+1}^{m+r_i} DCF_{i,n}$, is the value of the exposure at entry into the reference period for account i. Note that $EAD_{i,m}$ is equal to EAD_i, the EAD at the default date as in the Basel LGD calculations. The reference period is the area between the two blue dotted lines in Figure 1 and the length of the reference period is indicated by R (typically 12 to 24 months). The value r_i gives the distance between month on book m and the start of the recovery period. Account i enters the reference period at time $m+r_i$. Cashflows on account i occurring in the reference period are discounted to the time of default occurring on month on book m, i.e. $DCF_{i,n}=\frac{CF_{i,n}}{(1+e)^{n-m}}$. Cashflows, $CF_{i,n}$ are calculated as the difference between account balances now, versus account balances in the previous month, adding back the interest and the fees, subtracting the amount written off. Post write-off recoveries are added to the cashflows and are the additional expenses or recoveries after the write-off event (Witzany et al., 2012, pp. 8).

Similarly as in Joubert et al. (2018), a special dataset needs to be constructed. Each cashflow in the reference period will result in a separate record, with the unrecovered amount as a last record. Censoring is added to indicate if the account does not have information up to the end of the reference period.

The survival curve, $S(m,r,i)$, is then defined as the proportion of the exposure at entry into the reference period that remains in default from a specific time r $(m+r_i < r < m+r_i+R)$ until the end of the reference period, for account i that defaults at month on book m. The Kaplan-Meier estimate for this survival curve is the empirical value calculated from the data and is equal to:

$$\hat{S}\,(m,r,i)=\frac{\hat{DE}_{i,m}-\sum_{n=m+r_i+1}^{m+r_i+\mathrm{r}}\hat{DCF}_{i,n}}{\hat{DE}_{i,m}}.$$

The survival curve for the segment of account that defaulted on month on book m is then calculate as:

$$\hat{S_0}\,(m,r)=\frac{\hat{DE}_{m}-\sum_{n=m+r_i+1}^{m+r_i+\mathrm{r}}\hat{DCF}_{n}}{\hat{DE}_{m}},$$

where $\hat{DE}_m=\sum_{i\in I_m}\hat{DE}_{i,m}$ and $\hat{DCF}_n=\sum_{i\in I_m}\hat{DCF}_{i,n}$ with I_m indicating the accounts defaulting at month on book m.

The resulting Cox proportional hazards model is:

$$S\,(m,r,\boldsymbol{x}_\mathrm{i})=S_0\,(m,r)^{exp(\boldsymbol{x}_\mathrm{i}{}'\beta_m)},$$

where $S_0\,(m,r)$ is the baseline survival curve, i.e the Kaplan–Meier estimate of the base group. If an account falls outside of this base group, the baseline survival curve is shifted by $exp(\boldsymbol{x}_\mathrm{i}{}'\beta_m)$. The Cox model is fitted for each month on book segment.

The loss given default for account i that defaults at month on book m, is calculated as:

$$LGD_{i,m}=\frac{S\,(m,m{+}r_i{+}R,\boldsymbol{x}_\mathrm{i})}{S\,(m,m{+}r_i,\boldsymbol{x}_\mathrm{i})}.$$

The forward-looking LGD values ($LGD_{i,m+h}$) in the $ECL_{i,m}$ formula for each account i are then calculated using the above. These LGD values need to be adjusted for different macro-economic scenarios before using them.

2.2.4. Macro-economic model

The IFRS 9 accord (IFRS, 2014, pp. 189) requests banks to incorporate forward-looking macro-economic information into their estimation of lifetime expected credit losses. Various macro-economic scenarios are expressed in terms of macro-economic factors and probabilities are assigned to these scenarios. Therefore, the three components used in the calculation of the ECL need to be modelled with respect to the macro-economic factors. This adjustment is discussed below.

In order to adjust for different macro-economic scenarios, the $LGD_{i,m}$ values in the ECL calculation are expressed as a time series. Let $LGD_{c,i}$ be the LGD for account i, assuming the account defaults in calendar month c:

$$LGD_{c,i} = \frac{S\left(m_c, m_c + r_i + R, \boldsymbol{x}_{\mathrm{c,i}}\right)}{S\left(m_c, m_c + r_i, \boldsymbol{x}_{\mathrm{c,i}}\right)},$$

where $S\left(m_c, r, \boldsymbol{x}_{\mathrm{c,i}}\right) = S_0\left(m_c, r\right)^{exp\left(x'_{\mathrm{c,i}}\beta_{m_c}\right)}$ and m_c is the month on book for account i at calendar month c. Note $\boldsymbol{x}_{\mathrm{c,i}}$ is the set of covariates for account i at calendar month c and β_m as fitted before, i.e. the fitted model is applied to the portfolio on calendar months $c = 1,..,C$ to form a time series of LGD's.

For every calendar month c the LGD's are EAD averaged to obtain a portfolio LGD,

$$LGD_c = \frac{\sum\limits_i LGD_{c,i} \times EAD_{c,i}}{\sum\limits_i EAD_{c,i}},$$

where $EAD_{c,i}$ is the EAD for account i at calendar month c.

To incorporate the macro-economic factors into the time series, an 'Error Correction Model' (ECM) is used as, introduced by Engle and Granger (1987), using co-integration and error correction. Mohamed (2010) discusses and proposes implementation aspects of the ECM. If there exists a stationary linear combination between two non-stationary time series, the two variables combined are said to be 'co-integrated' (Granger, 1981) and (Granger, 1983).

The ECM may be implemented using four steps (Mohamed, 2010). The first step is to determine whether all the time series (LGD values and macro-economic variables) are integrated of the same order (Mohamed, 2009a). The second step is to demonstrate that the time series are co-integrated (Mohamed, 2009b). The third step is to generate residuals by regressing the LGD values on the macro-economic variables. The last step is to enter the lagged residuals from step three into a regression of the LGD differences on the macro-economic differences of the previous time period.

The augmented dicky-fuller (ADF) test is used in step one to test for stationarity (Mohamed, 2009a). To test for co-integration in the second step, an ordinary least squares regression is done between the LGD and the macro-economic variables. If the error terms from the regression is stationary, then co-integration exists (Mohamed, 2009b). The regression model in the third step can be defined as follows:

$$LGD_c = \alpha_0 + \alpha' \mathbf{z}_c + \varepsilon'_c$$

where $\mathbf{z}_c = \{z_1, \ldots, z_n\}$ is the n macro-economic variables, α_0 and $\alpha = \{\alpha_1, \ldots, \alpha_n\}$ are the parameters and ε'_c the residuals. The ECM model in step four is then expressed as:

$$\Delta LGD_c = \phi_0 + \phi' \Delta z_c + \phi_{n+1}\varepsilon'_{c-1} + \varepsilon_c$$

where $\Delta LGD_c = LGD_c - LGD_{c-1}$ and $\Delta z_c = \mathbf{z}_c - \mathbf{z}_{c-1}$. ϕ_0, $\phi = \{\phi_1, \ldots, \phi_n\}$, ϕ_{n+1} are the parameters, ε'_{c-1} is the lagged residual from step three and ε_c is the error term. Note that the term $\phi' \Delta z_c$ in the model may be extended to include different lags. Variable selection will be typically done using stepwise selection.

In general, the term $\varepsilon_c^{'}$ captures all other factors that influence the dependent variable LGD_c other than the independent variable $\mathbf{z}_c$ (this may be referred to as the 'long-term' error). In this sense, ECM can be seen as a way of combining the long run, a cointegrating relationship between the level variables and the short run relationship between the first differences of the variables (Mohamed, 2010).

Forecasted macro-economic values for $\mathbf{z}_c$ are obtained from Moody's analytics. For each macro-economic variable, a bull-, bear- and base case forecast is obtained. Let $\mathbf{z}_c^u$ be the optimistic scenario, $\mathbf{z}_c^b$ the base scenarios and $\mathbf{z}_c^d$ the pessimistic outcome. Given that actual data is used as far as it is known, the values of the three scenarios described above will be the same up until the point where the forecast starts. These variables are then used as inputs into the ECM to estimate LGD_c^u, LGD_c^d and LGD_c^b: an optimistic-, pessimistic- and base LGD by calender month. The average of each scenario's LGD values are calculated over the forecasted period. Two scalars are calculated as:

$$scalar_u = \frac{\sum_{c=1}^{f} LGD_c^u}{\sum_{c=1}^{f} LGD_c^b}$$

and

$$scalar_d = \frac{\sum_{c=1}^{f} LGD_c^d}{\sum_{c=1}^{f} LGD_c^b},$$

where f is the length of the forecast window.

A optimistic-, pessimistic- and base LGD by month on book are required for the ECL calculation. The above scalars are used to shift $LGD_{i,m}$ to an optimistic $LGD_{i,m}^u$ and a pessimistic $LGD_{i,m}^d$ where $LGD_{i,m}^u = LGD_{i,m} \times scalar_u$ and $LGD_{i,m}^d = LGD_{i,m} \times scalar_d$.

3. Data

In this section we will present the empirical LGD values and a description of the macro-economic variables that will be used in the ECM model.

3.1. Empirical LGD

The development dataset that is needed to model the IFRS 9 LGD includes the exposure at the entry into the reference period and values for all cashflows during the reference period. The default flag is needed to determine if an account is in default as at a specific month on book; the default flag and month on book indicator are stored on the development dataset. The account number, application date, closed date, effective interest rate and month are stored. The covariates, $\mathbf{x}$, that is used in the Cox proportional hazards model include behavioural-, customer-, and geographical- information and are added to the development dataset.

Data for an unsecured retail product from a South African bank are used in this paper. The development dataset that is used in this paper has 134741 defaulted accounts and the out of time dataset has 132642 accounts. Data from January 2005 up until December 2017 was considered for the development. The development reference period for the IFRS 9 LGD was selected as the most recent 24 month period available and the out of time the preceding 24 months. The empirical IFRS 9 LGD for the development- and out of time dataset are displayed in Figure 2. The empirical IFRS 9 LGD is calculated over a recent 24 month reference period, as was discussed in Section 2.2.2.

The exposure used for IFRS 9 is exposure at beginning of reference period, where as Basel uses exposure at default. IFRS9 cashflows are summed over the reference period, whereas Basel is summed over the workout period. The Basel LGD is therefore expected to be higher than the IFRS 9 LGD. The IFRS 9 LGD is calculated by subtracting the sum of the discounted cashflows over the reference period from the exposure at entry into the reference period, divided by the exposure at entry into the reference period. The Basel LGD is calculated by subtracting the sum of the discounted cashflows over the workout period from the exposure at the default point, divided by the exposure at the default point. The IFRS 9 LGD for accounts i that defaults at a specific month on book m is:

$$\hat{LGD}_{i,m} = \frac{\hat{DE}_{i,m} - \sum_{n=m+r_i+1}^{m+r_i+R} \hat{DCF}_{i,n}}{\hat{DE}_{i,m}},$$

where $\hat{DE}_{i,m} = \hat{EAD}_{i,m} - \sum_{n=m+1}^{m+r_i} \hat{DCF}_{i,n}$, is the value of the exposure at entry into the reference period for account i. Note that $\hat{EAD}_{i,m}$ is equal to $\hat{EAD}_i$, the EAD at the default date as in the Basel LGD calculations. The portfolio IFRS 9 LGD is calculated as:

$$\hat{LGD}_m = \frac{\hat{DE} - \hat{DCF}^*}{\hat{DE}},$$

where
$\hat{DE} = \sum_m \sum_i \hat{DE}_{i,m}$ and $\hat{DCF}^* = \sum_m \sum_i \sum_{n=m+r_i+1}^{m+r_i+R} \hat{DCF}_{i,n}$. The empirical Basel LGD is calculated over a workout period of 60 months and is expressed as:

$$\hat{LGD}_i = \frac{\hat{EAD}_i - \sum_{t=1}^{T_w} \hat{DCF}_{i,t}}{\hat{EAD}_i},$$

where $\hat{EAD}$ is the exposure at default for account i. The portfolio empirical Basel LGD is:

$$\hat{LGD} = \frac{\hat{EAD} - \hat{DCF}}{\hat{EAD}},$$

where $\hat{EAD} = \sum_i \hat{EAD}_i$ and $\hat{DCF} = \sum_i \sum_{t=1}^{T_w} \hat{DCF}_{i,t}$. The IFRS 9 development sample is constructed from the recent 24 month reference period and the out of time sample from the 24 month period preceding the development sample. The Basel development sample is constructed from the 24 month time period where benign economic conditions were prevalent. The out of time data is constructed from the 24 month time period preceding the development sample. The LGD values in all figures in this article are left out due to confidentiality.

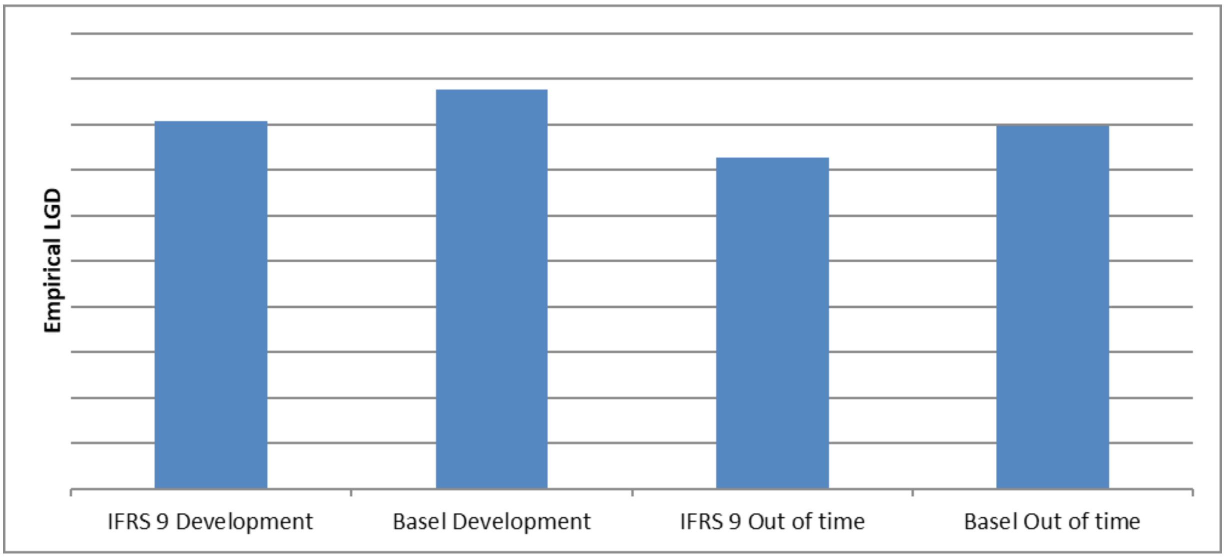


Figure 2: Empirical IFRS 9 LGD

3.2. Macro-economic variables

The following macro-economic variables were considered for the macro-economic model:

- The consumer price index (CPI) is the increase in the level of prices of a representative basket of goods purchased by consumers and households. This measures how much purchasing power in a country is eroded by price increases.
- The ratio of debt to disposable household income (DDHI) is a measure that indicates the ability of households to repay their debts. This measure is derived by dividing total monthly household debt by monthly income.
- The debt service ratio (DSR) is the proportion of household income that is spent on covering existing debt agreements.
- The M3 money supply is the money supply in circulation and indicates a country's liquid money supply.
- The gross domestic product (GDP) is an indication of the total local production of the economy.
- Nominal house price index (NHPI) is an index of the average house price level, without adjusting for inflation.
- The real house price index (RHPI) is an index measuring the average house price level, which adjusts for inflation.
- The prime interest rate is the rate at which the banks of South Africa lend money to customers.

- Debt affordability is the ratio of government debt relative to the resources available for repaying that debt.
- The leading indicator is a forecast of the the general health of the South African economy.
- Rand dollar exchange rate is the price of one dollar in rand terms.
- Liquidity spread is the premium that flows to a party willing to provide liquidity to a party that is demanding it.

The IFRS 9 LGD results are given in the following section.

4. Results

In this section we will present the results for the IFRS 9 LGD model as well as the results for the macro-economic model.

4.1. IFRS 9 LGD model

The IFRS 9 LGD is calculated by applying the methodology described in Section 2.2 to a retail credit portfolio of one of South Africa's major banks. The model was fitted to the development reference period for a portfolio and then applied to the same portfolio using both the development and out of time reference period. The empirical and estimated LGD values for each month on book are plotted in Figure 3. A difference in the overall level and directional movement of the development LGD and out of time LGD is observed. The reasons for these movements are provided in the discussion of Figure 4. A low number of accounts were observed where months on book is more than 180, causing the LGD values to be volatile past this point (not shown). The LGD value at months on book equal to 180 will therefore be used for accounts more than 180 months on book.

The development and out of time LGD values are compared in Figure 4. The development LGD was sorted in descending order and placed in deciles. The average out of time LGD vs. the average development LGD per decile is plotted. The development LGD values are lower than the out of time LGD values. The development samples reference period is the most recent 24 months and the out of times reference period is the preceding 24 months. The differences in strategies, customer behaviour and macro-economics are causing the LGD for these periods to be different. IFRS 9 LGD values are point in time in nature and movements in these curves over time are therefore anticipated. The point in time IFRS 9 LGD values will therefore be updated frequently to ensure that recent information is reflected in the LGD and ECL values. The movement between the development and out of time LGD values are therefore expected.

Figure 5 compares the estimated vs. the empirical IFRS 9 LGD values estimated over the development reference period as well as the estimated vs. the empirical IFRS 9 LGD values estimated over the out of time reference period . The estimated LGD was sorted in descending order and placed into deciles. The average estimated LGD vs. average empirical LGD per decile are plotted in Figure 5. Figure 5 shows that the estimated IFRS 9 LGD are close to the empirical IFRS 9 LGD values for both the development and out of time period and that the modelling methodology and adjustments made produce reasonable results.

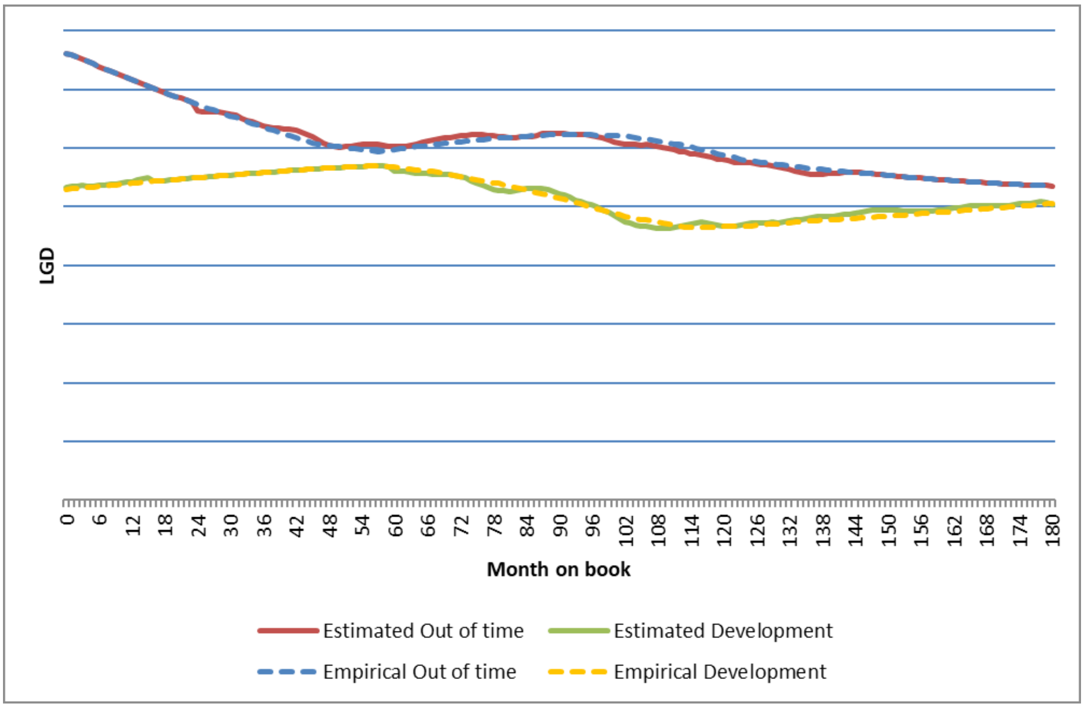


Figure 3: Empirical and estimated IFRS 9 LGD by month on book

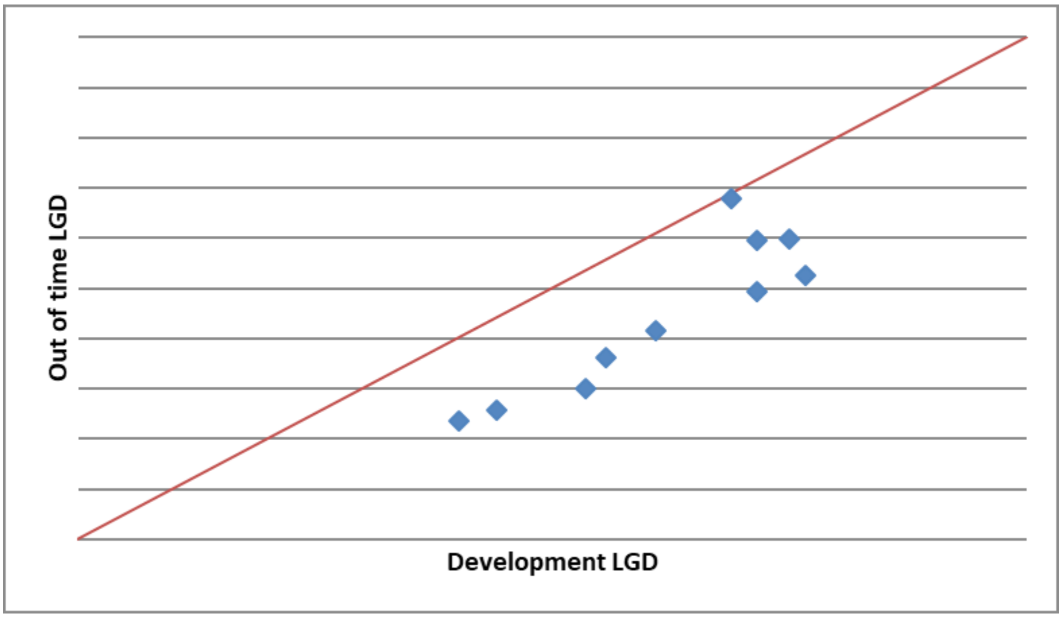


Figure 4: Development vs. out of time

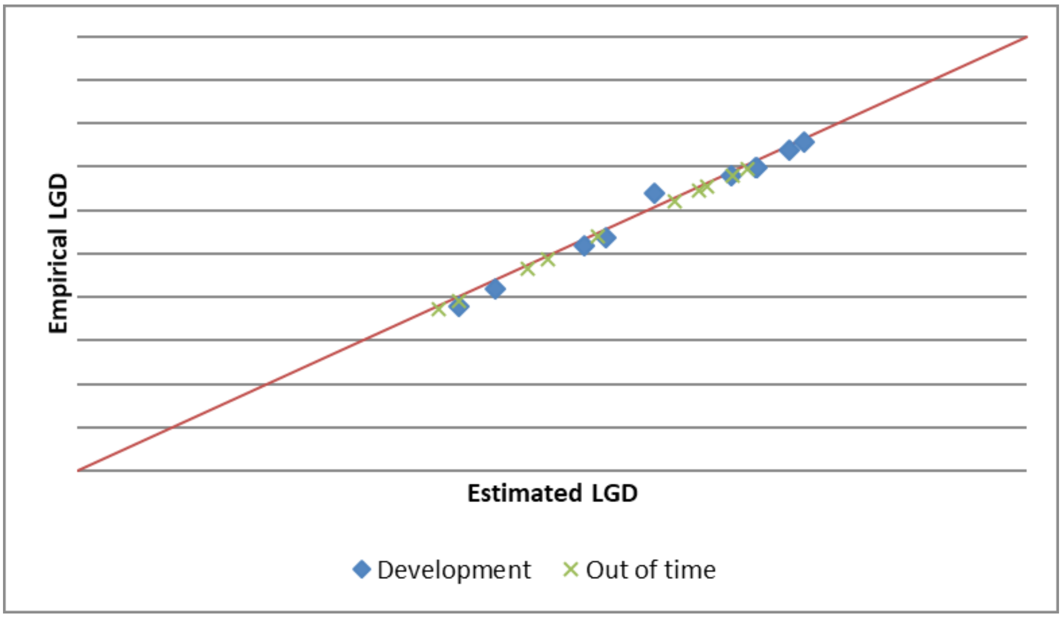


Figure 5: Accuracy graph

4.2. Macro-economic model (ECM)

The values for $LGD_{c,i}$ are displayed in Figure 6. A time series of macro-economic variables was fitted to the LGD curve in Figure 6 as the target in the ECM. Stationarity- and co-integration tests were performed, followed by the results for the ECM.

The first step of the ECM was performed to determine whether all the time series were integrated of the same order. The ADF test was used to test for stationarity. The regression procedure in SAS was used to perform the ADF test. The t-values from the regression procedure were compared to the t-statistic with a confidence level of 0.01. The t-statistic is equal to -3.524233 when a confidence level of 0.01 is considered. The t-values for the macro-economic variables are given in column two of Table 1. Each of these t-values are greater than the t-statistic value of -3.524233. We fail to reject the null hypothesis and conclude that the macro-economic variables are non-stationary when no differencing is performed. The first difference was taken and the ADF test applied to test if the first difference is stationary. The t-values for the differenced macro-economic variables are in column three of Table 1 and are less than the value of the t-statistic, -3.524233. These variables are therefore stationary and integrated of order one.

The second step was performed to demonstrate that the time series are co-integrated. An OLS regression with LGD_c as the target and the macro-economic variables as the independent variables was performed and the residuals stored. These residuals were then tested for stationarity by performing the ADF test. Column four of Table 1 contains the t-values for the errors. This ADF test showed that the error terms are stationary and conclude that LGD_c and the macro-economic factors are co-integrated.

The ECM model was fitted with $\Delta LGD_{c,i}$ as the target. The independent variables were the first

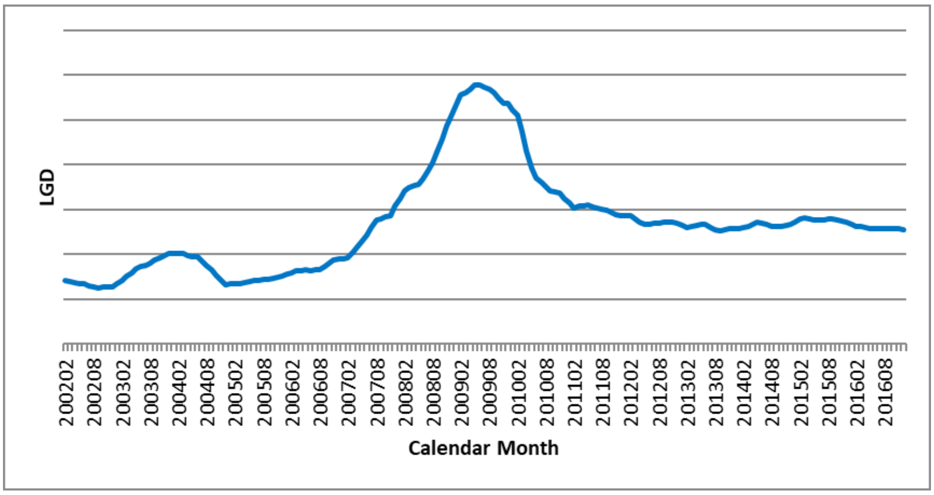


Figure 6: LGD by calendar month

Table 1: ADF test t-values

Macro-economic variable	t-value - no differencing	t-value - first difference	t-value - error
CPI	-1,74	-3,96	-4,29
DDHI	-2,23	-5,84	-5,78
debt affordability	-1,26	-5,75	-6,58
DSR	-1,22	-5,44	-3,97
GDP	-2,97	-4,24	-5,86
leading indicator	-0,44	-4,45	-5,48
liquidity spread	-0,72	-4,78	-4,51
M3 money supply	-0,71	-4,74	-4,85
NHPI	-1,92	-6,49	-4,82
prime interest rate	-1,87	-5,26	-5,36
rand dollar exchange rate	-2,32	-4,23	-4,34
RHPI	-0,93	-5,39	-5,51

difference of the macro-economic factors and the lagged error term from the previous regression. The macro-economic variables listed in Section 3.2 and their lags were considered for the long-term and short-term effect in the ECM. Lags 1 to 12 were taken for the first difference of each of the macro-economic factors and entered into a stepwise regression with an entry and exit significance level of 0.05.

The third step was to generate residuals done by regressing the LGD values on the macro-economic variables. The parameter estimates, lags and variance inflation factor for the variables are given in Table 2.

Table 2: Results for the short-term portion of the ECM

Parameter estimates	$\phi_0 = -0,32$	$\phi_1 = 0,41$	$\phi_2 = -0,51$	$\phi_3 = -0,33$
Variable	Intercept	GDP	DDHI	Prime
Lag used	-	Lag 3	Lag 8	Lag 7
VIF		2,44	1,02	2,42

The last step was to enter the lagged residuals from step three into a regression of the LGD differences on the macro-economic differences of the previous time period. The parameter estimates, lags and variance inflation factor for the variables are given in Table 3.

Table 3: Results for the long-run portion of the ECM

Parameter estimates	$\alpha_0 = 0$	$\alpha_1 = 0,04$	$\alpha_2 = -0,03$
Variable	Intercept	Leading indicator	Debt affordability
Lag used	-	Lag 1	Lag 5
VIF		1,37	1,59

4.3. Macro-economic scenarios

The fitted ECM model is applied to the forecasted covariates of the optimistic-, pessimistic- and base macro-economic scenarios. The covariates for these scenarios are $\mathbf{z}_c^u$, $\mathbf{z}_c^d$ and $\mathbf{z}_c^b$ and the resulting LGD by calendar months are LGD_c^u, LGD_c^d and LGD_c^b, respectively. These LGD values are displayed in Figure 7.

The values for $scalar_u$ and $scalar_d$ were calculated from the forecasted portions of the LGD values that are displayed in Figure 7. The LGD by month on book values for each of the three macro-economic scenarios were calculated as $LGD_{i,m}^u = LGD_{i,m} \times scalar_u$ and $LGD_{i,m}^d = LGD_{i,m} \times scalar_d$.

These LGD values together with a macro-economic adjusted PD and EAD were entered into the ECL formula to calculate an ECL value for each scenario. A marginal PD model using a frequentist approach was used, with similar macro-economic adjustment as in the LGD model. The same set of macro-economic variables are considered and forecasts are obtained from Moodys analytics. The

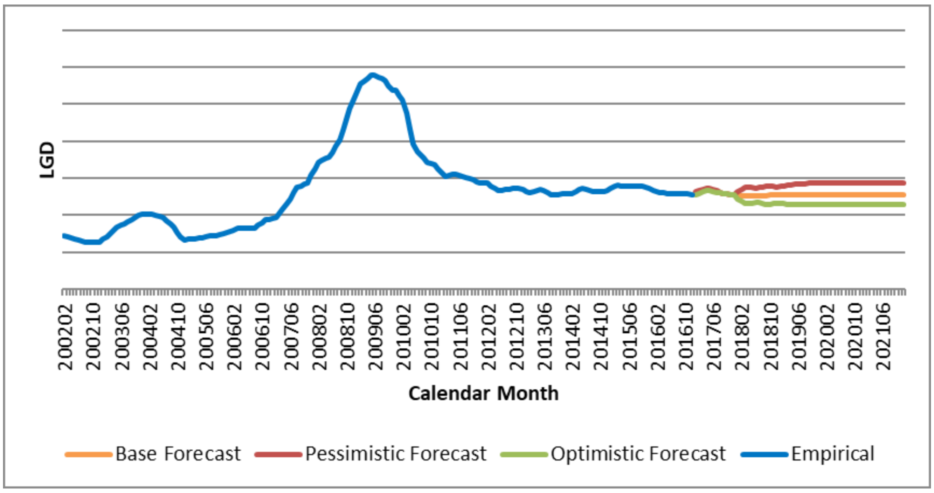


Figure 7: LGD by calendar month

EAD is estimated using the amortization schedule of each account with discounting. The probability of each scenario occurring are provided and are used to calculate a weighted ECL value. The percentage increase in the macro-economic adjusted IFRS LGD, percentage ECL increase and scenario weights for each of the macro-economic scenarios are given in Table 4. An LGD increase of 5.17% and a PD increase of 3.29% is observed for the pessimistic scenario. These increases results in an ECL increase of 4.39% for the pessimistic scenario. The LGD decreases by 4.03% and the PD decrease by 2.07% for the optimistic scenario, resulting in an ECL decrease of 3.27%. The GDP, Prime and Debt affordability macro-economic variables entered into the PD and LGD models.

Table 4: Macro-economic adjusted IFRS 9 LGD values

	Base scenario	**Optimistic scenario**	**Pessimistic Scenario**
% change in IFRS 9 LGD	0%	-4.03%	5.17%
% change in IFRS 9 PD	0%	-2.07%	3.29%
Probability of scenario occurring	40%	30%	30%
% change in ECL	0%	-3.27%	4.39%

The combined scenario weighted *%ECL* increase is 0.3%. A conclusion of this paper now follows.

5. Conclusion

Survival analysis is one of several methods used to predict loss given default (LGD) for Basel regulatory purposes. When using survival analysis to model LGD, one proposed methodology is the default weighted survival analysis (DWSA) method. This paper is aimed at adapting the DWSA method used to model the Basel LGD to estimate the LGD for IFRS 9 impairment requirements. This IFRS 9 LGD is used in the calculation of the ECL as per the IFRS 9 accord. The IFRS 9 LGD methodology that is described in this paper makes use of survival analysis to estimate the LGD. The Cox proportional hazards model allows for a baseline survival curve to be adjusted to produce survival curves for different segments of the portfolio. The DWSA methodology allows for over recoveries, default weighting and negative cashflows. The forward-looking LGD values are adjusted for macro-economic scenarios. An ECM model was used to predict a macro-economic model. An ECL is calculated for different macro-economic scenarios. These ECL values are probability weighted to produce a single ECL number.

The survival analysis methodology to produce the IFRS 9 LGD was validated and tested on a South African retail bank dataset for accuracy by comparing the empirical and estimated LGD by deciles. All the necessary stationarity and co-integrations tests were performed when fitting the ECM.

Future research could focus on the IFRS 9 probability of default (PD) component that is used to calculate the expected credit loss. Basel PD modelling methodologies could be adapted for IFRS 9. Survival analysis, machine learning and behavioural scorecard techniques can be considered to model the IFRS 9 PD.

References

APTIVAA (2016a). Building blocks of impairment modeling (issue 02). *URL:https://www.slideshare.net/DianaP14/blog-02-building-blocks-of-impairment-modeling*.

APTIVAA (2016b). Cash shortfall and lgd two sides of the same coin. *URL:http://www.aptivaa.com/blog/cash-shortfall-lgd-two-sides-of-the-same-coin/*.

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2015). Guidance on accounting for expected credit losses. *URL:https://www.bis.org/bcbs/publ/d311.pdf/*.

BASTOS, J. (2010). Forecasting bank loans loss given default. *Journal of banking and Finance, 34. 2510-2517*.

BEERBAUM, D. (2015). Significant increase in credit risk according to ifrs 9: Implications for financial institutions. *Int J Econ Manag Sci 4:287. doi: 10.4172/21626359.1000287*.

BLACK, B., LEVINE, G., AND LICARI, J. (2015). Probability-weighted outcomes under ifrs 9 - a macroeconomic approach. *Moody's Analytics, Risk Perspective, The Convergence of Risk, Finance and Accounting, Volume VII*.

BRAUN, C. (2004). The prediction error of the chain ladder method applied to correlated run-off triangles. *Astin Bulletin, Vol. 34, No. 2, 2004, pp. 399-423*.

BROWN, I. (2014). Developing credit risk models using sas enterprise miner and sas/stat: Theory and application. *Cary, NC:SAS Institute inc*.

CHAWLA, G., FOREST, L., AND AGUAIS, S. (2016). Point-in-time (pit) lgd and ead models for ifrs9, cecl and stress testing. *URL:http://www.henrystewartpublications.com/jrm/*.

ENGLE, R. AND GRANGER, C. (1987). Cointegration and error-correction: Representation, estimation, and testing. *Econometrica 55 March, pp. 251-276*.

EUROPEAN BANKING AUTHORITY (EBA) (2016). Consultation paper eba/cp/2016/10: Draft guidelines on credit institutions: credit risk management practices and accounting for expected credit losses. *URL: https://www.eba.europa.eu/documents/10180/1532063/EBA-CP-2016-10+%28CP+on+Guidelines+on+Accounting+for+Expected+Credit%29.pdf*.

GLOBAL PUBLIC POLICY COMMITTEE (GPPC) (2016). The implementation of ifrs 9 impairment requirements by banks: Considerations for those charged with governance of systemically important banks. *URL:http://www.ey.com/Publication/vwLUAssets/Implementation_of_IFRS_9_impairment_requirements_by_systemically_important_banks/$File/BCM-FIImpair-GPPC-June2016%20int.pdf*.

GRANGER, C. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics, 16, 121-30*.

GRANGER, C. (1983). Co-integrated variables and error-correcting models. *University of California, San Diego, Department of Economics Working Paper: 83-13*.

IFRS (2014). Irfs9 financial instruments: Project summary. *URL:http://www.ifrs.org/Current-Projects/IASB-Projects/Financial-Instruments-A-Replacement-of-IAS-39-Financial-Instruments-Recognitio/Documents/IFRS-9-Project-Summary-July-2014.pdf*.

JOUBERT, M., VERSTER, T., AND RAUBENHEIMER, H. (2018). Default weighted survival analysis to directly model loss given default. *Accepted for publicatiin in the south african statistival journal (sajs)*.

MIU, P. AND OZDEMIR, B. (2017). Adapting the basel ii advanced internal ratings based models for international financial reporting standard 9. *Journal of Credit Risk 13(2)*.

MOHAMED, I. (2009a). Simulating time series analysis using sas part i the augmented dicky- fuller (adf) test. *L3 Communications-ETIS, Reston, VA*.

MOHAMED, I. (2009b). Simulating time series analysis using sas part ii co-integration. *L3 Communications-ETIS, Reston, VA*.

MOHAMED, I. (2010). Simulating time series analysis using sas part iii error correction model (ecm). *L3 Communications-ETIS, Reston, VA*.

WITZANY, J., RYCHNOVSKY, M., AND CHARAMZA, P. (2012). Survival analysis in lgd modelling. *European Financial and Accounting Journal, 2012, vol. 7, no. 1, pp. 6-27*.

Chapter 5

Conclusion

Contents of Chapter 5

List of Figures

List of Tables

CONCLUSION

This chapter provides the concluding remarks on the thesis as a whole and suggests future research ideas.

1. Summary

This thesis focused on developing LGD models suitable for the Basel Accord and the IFRS 9 standard. LGD models were introduced to model the Basel LGD directly and indirectly for secured and unsecured portfolios, respectively. Furthermore, the Basel LGD was adjusted to model the IFRS 9 LGD.

Chapter 2 introduced improvements to the direct approach to modelling the LGD for unsecured retail portfolios. The modelling of LGD is brought into closer alignment with regulations by introducing the default weighted survival analysis (DWSA) method. The methodology was also adapted to allow for negative cashflows and for over recoveries that can occur in practice. The newly introduced technique improves on existing LGD modelling techniques.

Chapter 3 focused on the indirect approach to model LGD for a secured retail portfolio. The chapter specifically investigated whether the predictive power of the model could be improved if the traditionally used logistic regression was replaced by survival analysis in order to model the probability component. Survival analysis is well suited to model these components since it allows for censoring. By using censoring, incomplete accounts are modelled more accurately. Also, survival analysis allows for an outcome period that varies in length. This aids in the accuracy of the prediction of the three probability components (write-off, cure and incomplete).

In Chapter 4, the DWSA method was adapted to model the IFRS 9 LGD. A baseline survival curve was adjusted to produce survival curves for various segments according to the Cox proportional hazards model. Using the DWSA method, over recoveries and negative cashflows are allowed for. Macro-economic variables were modelled onto the LGD using an error correction model (ECM). The LGD estimates were then adjusted for different economic scenarios. In each scenario an ECL was calculated and these values were probability weighted to produce a single ECL value.

2. Key findings

The key findings of each chapter will now be discussed, followed by more general findings across all chapters.

2.1. Direct Basel LGD model development approach (Chapter 2)

The EWSA approach to make use of survival analysis to model LGD directly existed in literature (Witzany, Rychnovsky and Charamza, 2012), but certain adaptations were required to align this model to regulations. The Basel Accord stipulates that LGD values are required to be default weighted and not exposure weighted as was the case in the above-mentioned article. The survival analysis technique was adapted to cater for default weighting and the DWSA approach was introduced. Negative cashflows were excluded from the study by Witzany et al. (2012), but existed in the data of the unsecured retail credit product under question. The LGD model development methodology was adapted to cater for negative cashflows by developing two separate survival curves; one for positive cashflows and one for negative cashflows, and then combining them to produce a single LGD value. Another occurrence on the data of the unsecured product, that was not catered for in the article by Witzany et al. (2012), was that of over recoveries. Over recoveries occur when the sum of the discounted recovery stream exceeds the exposure at the default point. Traditional survival curves do not cater for this situation and an adaptation was made to allow for over recoveries into the calculation of LGD.

The MSE, bias and variance for the retail bank LGD and the simulated LGD are displayed in Figure 1 and Figure 2, respectively. The DWSA approach yielded the best results for both the retail bank LGD and the simulated LGD. The model resulted in the lowest MSE, lowest bias and the lowest variance for both the retail bank LGD and the simulated LGD.

The DWSA approach gave the best results when compared to other methodologies. In the next section, an indirect Basel LGD model development technique will be described.

In Chapter 2, the DWSA approach gave the best results when compared to other methodologies and allows for the following enhancements:

- Default weighting of the Basel LGD estimate
- The inclusion of negative cashflows into the modelling approach
- Catering for over recovery

2.2. Indirect Basel LGD model development approach (Chapter 3)

The indirect LGD model consists of two components; the loss severity component and the probability component. The loss severity components were calculated as was done by Leow and Mues (2012). Survival analysis was used to model the probability of cure, probability of write-off and probability of remaining incomplete since logistic regression only allows the modelling of one probability (the probability of write-off and not write-off). A benefit when using survival analysis (rather than logistic regression) is the inclusion of incomplete accounts by using censoring. The importance

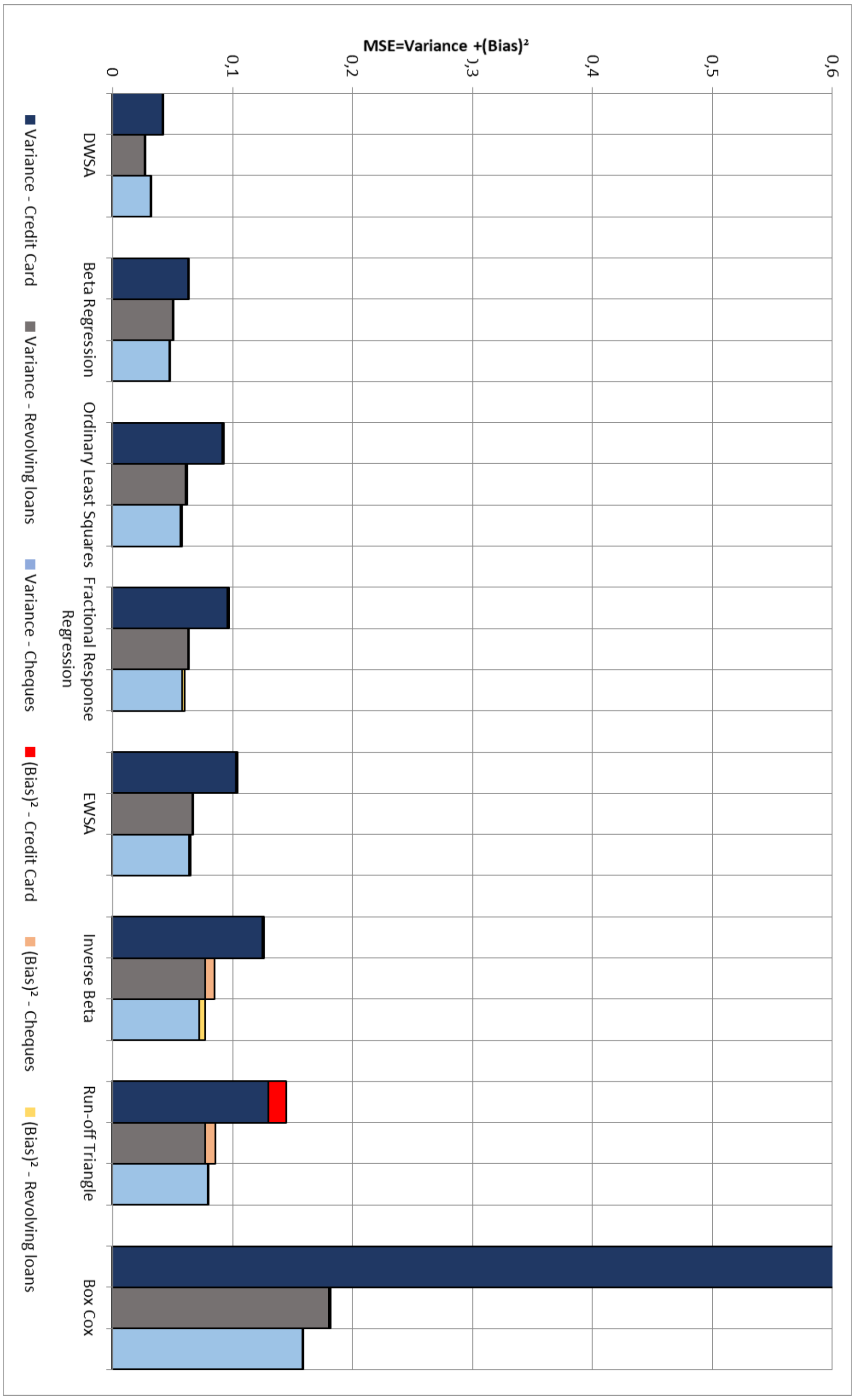


Figure 1: Retail bank dataset results for direct modelling approaches

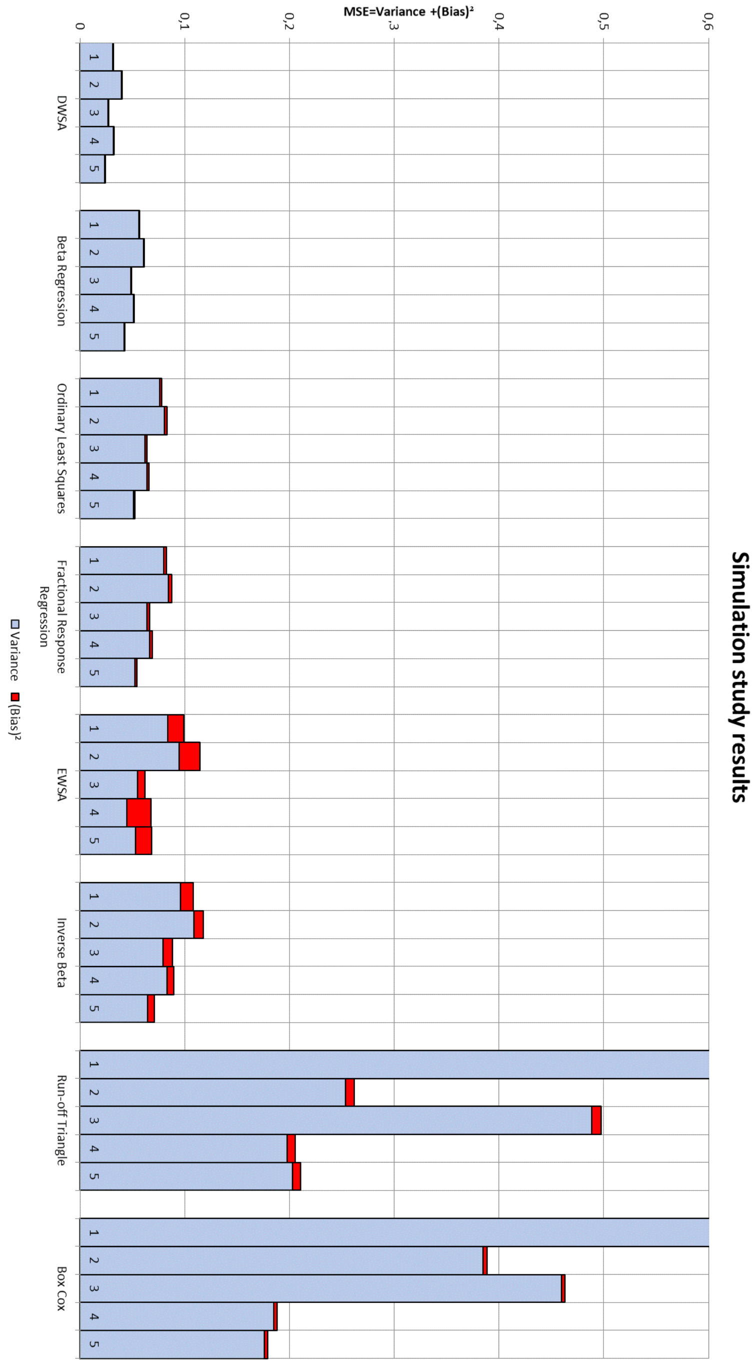


Figure 2: Simulation study results for direct modelling approaches

of including incomplete information was known to the author due to industry experience. This inclusion will cause the Basel LGD to be more accurate. The challenge of introducing incomplete accounts were two-fold. Firstly, the nature of incomplete accounts is such that they only have information up to a certain point and no further information after that point, therefore an appropriate technique is required to deal with these accounts. Secondly, the target will not be binary anymore and the use of a binary logistic regression will become imperfect. Survival analysis naturally lends itself to solving the first challenge; incomplete accounts are censored when no further information is available and these accounts are correctly accounted for in the estimation of LGD. The second challenge was solved by introducing the cumulative incidence function into this competing risk environment.

The MSE, bias and variance for the retail bank LGD and the simulated LGD are displayed in Figure 3 and Figure 4, respectively. The survival analysis approach resulted in the lowest MSE, lowest bias and the lowest variance on both the retail bank and the simulated data compared to the logistic regression approach.

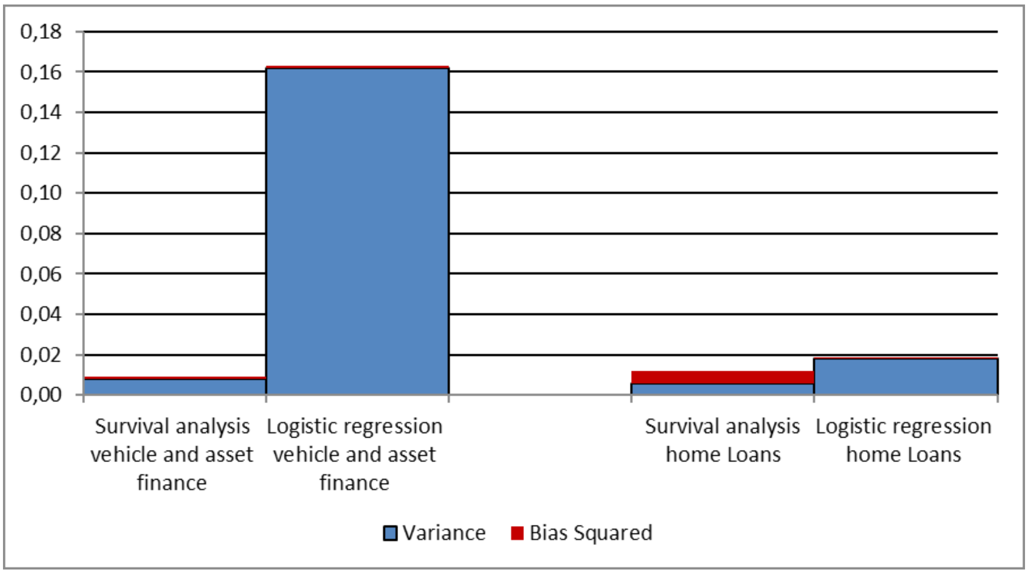


Figure 3: Retail data MSE, variance and bias

The corresponding values are given in Table 1.

	Vehicle and Asset finance			Home loans portfolio		
Method	Variance	MSE	Bias	Variance	MSE	Bias
Survival	0.008	0.009	-0.037	0.006	0.012	0.077
Logistic	0.162	0.163	-0.082	0.018	0.019	-0.031

Table 1: Retail data

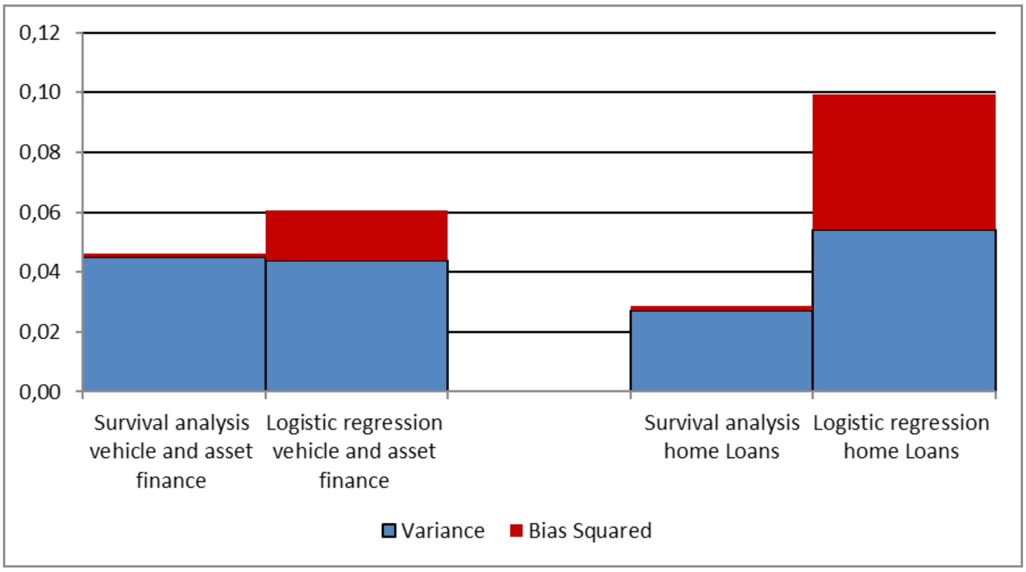


Figure 4: Simulated MSE, variance and bias

The corresponding values are given in Table 2.

	Vehicle and Asset finance			Home loans portfolio		
Method	Variance	MSE	Bias	Variance	MSE	Bias
Survival	0.045	0.046	0.034	0.027	0.029	0.04
Logistic	0.044	0.061	-0.129	0.054	0.1	0.213

Table 2: Simulated data

Survival analysis gave a better result than when logistic regression was used to model the probability component. The DWSA approach was used in Chapter 2 to model the Basel LGD and will be adopted in the next section to model the IFRS 9 LGD.

In Chapter 3, survival analysis gave a better result than when logistic regression was used to model the probability component. There are four primary enhancements by substituting logistic regression with survival analysis. Firstly, survival analysis includes incomplete accounts in the model via censoring. Secondly, the information content of the model is improved by including survival time. The third improvement to the model gained by including survival analysis is that a time-varying outcome window instead of a fixed outcome window is achieved. This aligns the model to reality. The fourth benefit is that model accuracy is improved by including more than two outcomes in the target variable of the probability component.

2.3. IFRS 9 LGD model development approach (Chapter 4)

The Basel and IFRS 9 LGD are based on the same principles (for IFRS 9 Stage 1 the same loss distribution is used as Basel) and it is therefore natural to adopt the Basel LGD to model the IFRS 9 LGD. The DWSA approach was used in Chapter 2 to model the Basel LGD and adapted in Chapter

4 to model the IFRS 9 LGD. The IFRS 9 LGD was segmented by month on book and the LGD was calculated for each month on book by applying the DWSA approach. The forward-looking LGD for an account at a specific point in time can be read off an LGD by month on book table. The IFRS 9 standard further required the forward-looking IFRS LGD to be adjusted for various macro-economic scenarios. Macro-economic variables were regressed onto a time series of LGD by calendar month in the form of an ECM model. An optimistic, pessimistic and base scenario were entered into the fitted ECM model. These scenarios were expressed in terms of the macro-economic variables in the fitted ECM model. A time series of LGD by calendar month was used to calculate the average estimate future LGD for each scenario. The ratio of the optimistic and pessimistic scenarios was taken to the base scenario, respectively. These two ratios were applied to the base LGD by month on book curve to produce an optimistic and pessimistic LGD by month on book curve.

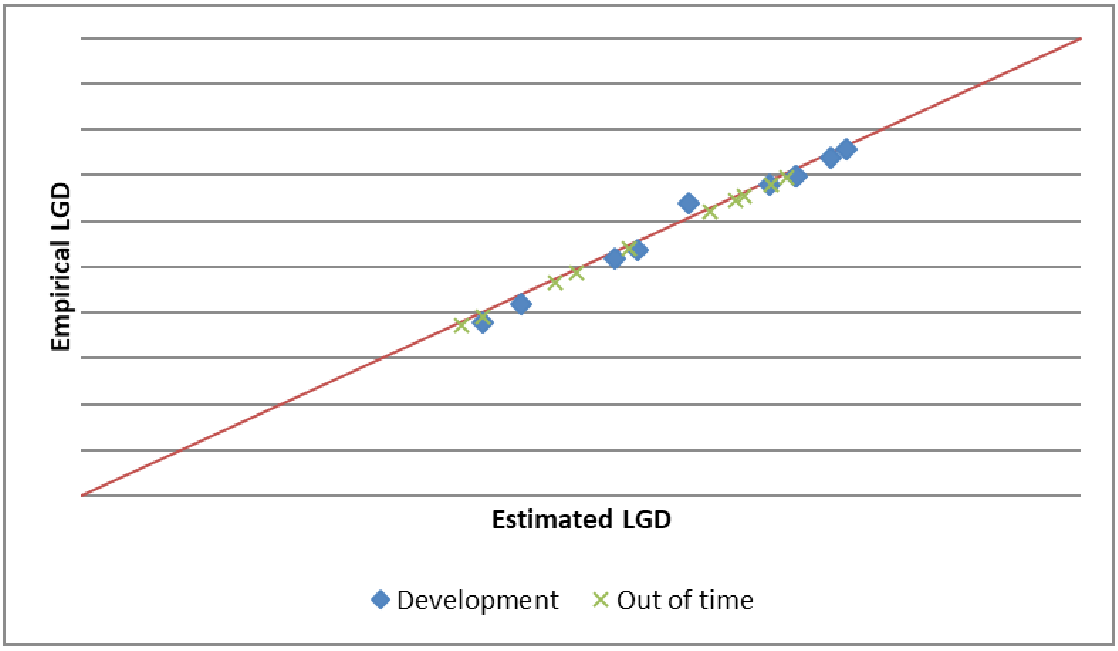


Figure 5: Accuracy graph

Figure 5 shows that the estimated IFRS 9 LGD follow the empirical IFRS 9 LGD closely on both the development and out of time datasets. The 45 degree red line indicates where the estimated IFRS 9 LGD is equal to the empirical IFRS 9 LGD.

The LGD by month on book values for each of the three macro-economic scenarios were calculated. PD and EAD were adjusted with macro-economic variables, alongside the above-mentioned LGD values and were included in the ECL formula to calculate the ECL for each scenario. The weighted ECL value was derived by weighting each outcome by its probability of occurring. Table 3 sets out the percentage increase in the IFRS 9 LGD that is macro-economically adjusted, the percentage increase in ECL and the weightings per scenario. An LGD increase of 5.17% and a PD increase of 3.29% is observed for the pessimistic scenario. These increases results in an ECL increase of 4.39% for the pessimistic scenario. The LGD decreases by 4.03% and the PD decrease by 2.07% for the optimistic scenario, resulting in an ECL decrease of 3.27%.

Table 3: Macro-economic adjusted IFRS 9 LGD values

	Base scenario	Optimistic scenario	Pessimistic Scenario
% change in IFRS 9 LGD	0%	-4.03%	5.17%
% change in IFRS 9 PD	0%	-2.07%	3.29%
Probability of scenario occurring	40%	30%	30%
% change in ECL	0%	-3.27%	4.39%

The IFRS 9 LGD for impairment requirements was modelled in Chapter 4. The IFRS 9 LGD was adapted from the DWSA methodology, which allows for over recoveries, default weighting and negative cashflows. This IFRS 9 LGD was used in the calculation of the ECL as per the IFRS 9 Accord and the forward-looking LGD values were adjusted for macro-economic scenarios. An ECL was calculated for different macro-economic scenarios and the ECL values were probability weighted to produce a single ECL.

A summary the Cox proportional hazards models, that is used in Chapters 2 to 4, is given in the next section.

2.4. Cox proportional hazards models (Chapters 2 to 4)

The Cox proportional hazards model was used in the development of the direct Basel LGD, indirect Basel LGD and the IFRS 9 LGD model. The Cox proportional hazards model allows for covariates into the model when fitting survival curves. This is advantageous in the sense that a population does not have to be separated into various segments upfront and a model fitted to each segment; a base curve for a segment is calculated and the model adjusted this base curve to be representative of the other segments. Up from segmentation can be greatly reduced, with the possible value that segmentation might add remaining. A stepwise regression is used to select the variables into the Cox proportional hazards model. The combination of the attributes of these variables can be seen as the segmentation scheme used in the model. The stepwise regression makes use of statistical testing to select variables and the segmentation scheme is therefore selected in a statistical manner. Although statistical testing is considered when upfront segmentation is used, human bias may occur when deciding on upfront segments. Another important fact to note is that the attributes for each segment are grouped in such a way that the hazard rate makes logical sense and that the hazard rate separates well between attributes. This ensures good separation for segments. The Basel Accord requires pooled estimates to be used and this approach will assess the pooling of accounts into segments.

3. Future research

The topic of discussion in this thesis was mainly the LGD component for both Basel and IFRS 9. Literature exists for the Basel PD component, but is lacking for the IFRS 9 PD. The reason for this is the recent change from IAS 39 to IFRS 9. As with the LGD component, the Basel PD modelling methodologies could be adapted for IFRS 9; reason being that these components come from the same distribution (for stage 1). Survival analysis, machine learning and behavioural scorecard techniques can be considered to model the IFRS 9 PD. The PD model used in Chapter 4 follows a frequentist approach (i.e. empirical PD) to estimate the IFRS 9 PD. This approach can be used to benchmark other PD techniques.

According to the IFRS 9 standard, accounts that have significantly increased in credit risk (SICR) since origination should be placed into Stage 2. The assumption is that all banks store their account origination information and will be able to compare the current performance of an account against origination information. Not all banks have the appropriate origination information stored away, and research can be conducted to produce an approach that can be used to identify accounts that have significantly increase in credit risk since origination, given the constraint of lacking origination data.

In addition, the generalized additive proportional hazard model (Hastie and Tibshirani, 1990, pp. 211–218) may be used as an alternative to the proportional hazards model in the DWSA model. The additive model will allow for nonlinear covariate effects. Additional topics of further research can be to use B splines (Ohlsson and Johansson, 2010, pp. 106–108) as the smoothing function for each of the covariates fitted in the DWSA model.

Basel suggests an ASRF model for calculating RWA, but does not limit banks to make use of this framework. Further research can be conducted where other models are considered to calculate RWA.

References

HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized Additive Models. Chapman and Hall/CRC.*

LEOW, M. AND MUES, C. (2012). Predicting loss given default (lgd) for residential mortgage loans: A two-stage model and empirical evidence for uk bank data. *International Journal of Forecasting, pp 183195.*

OHLSSON, E. AND JOHANSSON, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models.*

WITZANY, J., RYCHNOVSKY, M., AND CHARAMZA, P. (2012). Survival analysis in lgd modelling. *European Financial and Accounting Journal, 2012, vol. 7, no. 1, pp. 6-27.*