Volume 34 (2), pp. 107–132
http://orion.journals.ac.za

ORiON
ISSN 0259–191X (print)
ISSN 2224–0004 (online)
©2018

Making use of survival analysis to indirectly model loss given default

M Joubert[*] T Verster[†] H Raubenheimer[†]

Received: 23 February 2018; Revised: 16 May 2018; Accepted: 10 July 2018

Abstract

A direct or indirect modelling methodology can be used to predict *Loss Given Default* (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 mean squared error, bias and variance for the two methodologies were compared and it was shown that the use of survival analysis 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 according to the results obtained it 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

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 methodologies. A subjective LGD methodology makes use of expert judgement and is used for low default portfolios, portfolios with insufficient data and new portfolios. Objective LGD methodologies can be classified into explicit

[*] Quantitative model development, Investec, Sandton, South Africa, email: joubertmorne9@gmail.com

[†] Centre for Business Mathematics and Informatics, North-West University, South Africa, email: Tanja.Verster@nwu.ac.za, Helgard.Raubenheimer@nwu.ac.za

http://dx.doi.org/10.5784/34-2-588

and implicit methodologies. An explicit methodology allows for the direct computation of LGD, whereas with an 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 methodologies and expert judgement is categorized as a subjective method [8, 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 [2, 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 [6, 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 [8, p.157-158]. Figure 1 contains a diagram that illustrates the classification of the various LGD approaches.

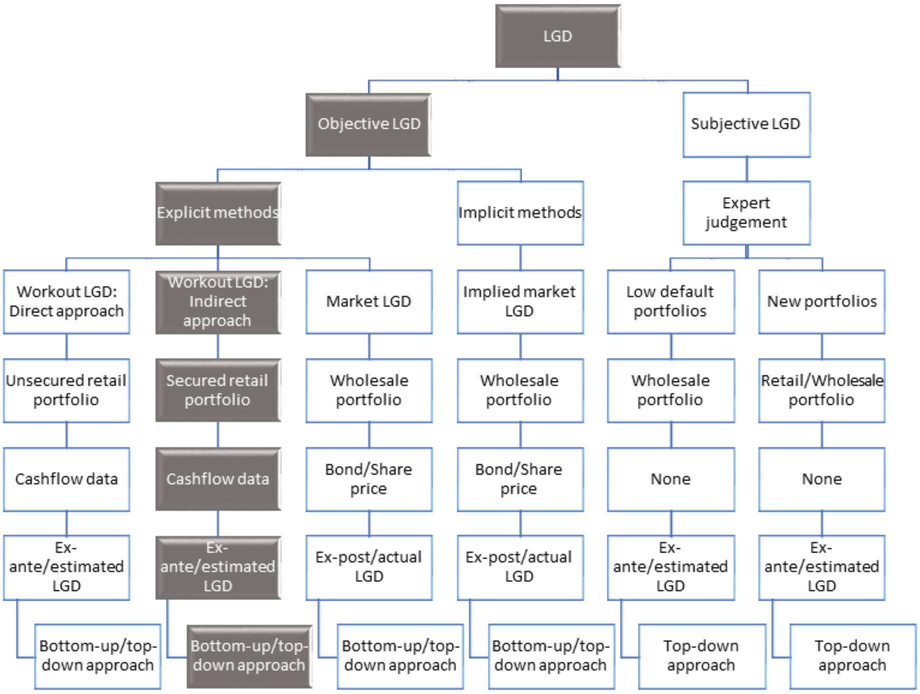

Figure 1: LGD approaches classified.

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 [14, 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 [11]. The papers by Bellotti & Crook [4], Lotheram *et al.* [12] and Qi & Zhao [17] 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 [19, p.548]. PD models have been used in the retail sector since 1960 [1, Section xxiv]. With the advent of regulatory capital changes driven by legislation such as the Basel Capital Accord [3], more focus and emphasis has been placed on LGD methodologies. 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 [8, p.158].

Zhang & Thomas [21], Schmidt [15] and Witzany *et al.* [20] made use of a linear regression, run-off triangle and Cox proportional hazards regression, respectively, to model the recovery rate directly. Tong *et al.* [19] made use of a zero adjusted gamma model to model the LGD directly. The LGD was modelled indirectly by Somers & Whittaker [16], Qi & Yang [14], Zhang *et al.* [22], Tong *et al.* [18] and Leow & Mues [11].

Leow & Mues [11] indirectly modelled LGD by modelling a probability component and a 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 [7, p.34] mentions that incomplete accounts carry valuable information and excluding this information will lead to the underestimation of LGD.

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 times 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, as well as the probability component and the loss severity component. 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 [2].

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

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

In [11], the probability components, $P_{i,t}\,(W_\tau)$, $P_{i,t}\,(C_\tau)$ and $P_{i,t}\,(I_\tau)$ are predicted 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 [11] and how the haircut model is used to predict the loss severity component.

3 Modelling methodology

The approach in [11] 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. The probability of write-off, $P_{i,t}\left(W_{\tau}\right)$, was modelled in [11] by making use of logistic regression

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

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

In this section, a discussion on how survival analysis instead of logistic regression can be used to predict the probability components $P_{i,t}\left(W_{\tau}\right)$, $P_{i,t}\left(C_{\tau}\right)$ and $P_{i,t}\left(I_{\tau}\right)$. This is followed by a description of the haircut model that will be used. No changes are made to the haircut model that [11] 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 $P_{i,t}\left(W_{\tau}\right)$, $P_{i,t}\left(C_{\tau}\right)$ and $P_{i,t}\left(I_{\tau}\right)$, 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. In [20] the survival function, S, is defined as

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

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, $h\left(t\right)=\frac{f(t)}{S(t)}$, is the instantaneous rate of exit at t, given that survival has been attained up to point t. The survival function, $S\left(t\right)=e^{-H(t)}$, is expressed in terms of the cumulative hazard function $H\left(t\right)=\int_0^t\lambda\left(t\right)ds$ [20]. 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 this 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, $S^w\left(t\right)$, and a survival function for cure, $S^c\left(t\right)$, 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

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

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

$S^{w}\left(t\right)$ 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 $S^{w}\left(t\right)=P(T>t)$, with not write-off as the event. Similarly, $S^{c}\left(t\right)$ 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 $S^{c}\left(t\right)=P(T>t)$, with not curing as the event.

The two survival functions, $S^{w}\left(t\right)$ and $S^{c}\left(t\right)$, 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 [10].

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

$$S\left(t\right)=\left[S_{0}(t)\right]^{\exp(\mathbf{x}'\beta)}.$$

The formula states that the survival curve at time t is a function of two quantities. The first of these, $S_{0}\left(t\right)$, 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, $S_{0}\left(t\right)$, 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, $S_{0}\left(t\right)$, will be adjusted,

$$S_{0}\left(t\right)^{\exp(\mathbf{x}'\beta)},$$

to yield a survival curve, $S\left(t\right)$, 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 [20] as

$$h(t,x)\;=\;h_{0}\left(t\right)\exp(\mathbf{x}'\beta),$$

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

$$S\left(t,\mathbf{x}\right)=\exp\left(-\int_{0}^{t}h_{0}\left(s\right)\exp\left(\mathbf{x}^{'}\beta\right)\right)=S_{0}\left(t\right)^{\exp\left(\mathbf{x}^{'}\beta\right)},$$

where $S_{0}\left(t\right)=\exp\left(-\int_{0}^{t}h_{0}\left(s\right)\right)$. 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

$$L_{i}\left(\beta\right)=\frac{h(t,\;x_{\mathbf{i}})}{\sum\limits_{j\varepsilon A_{i}}h(t,\;\mathbf{x_{j}})}=\frac{\exp(-\mathbf{x_{i}}'\beta)}{\sum\limits_{j\varepsilon A_{i}}\exp(-\mathbf{x_{j}}'\beta)}$$

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 maximised 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 maximised. 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. [20] 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 $[t,\tau]$

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

3.2 Advantages of using survival analysis

The advantages of using the Cox proportional hazards approach instead of a logistic model [10] 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 [11]. 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 [11], 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:

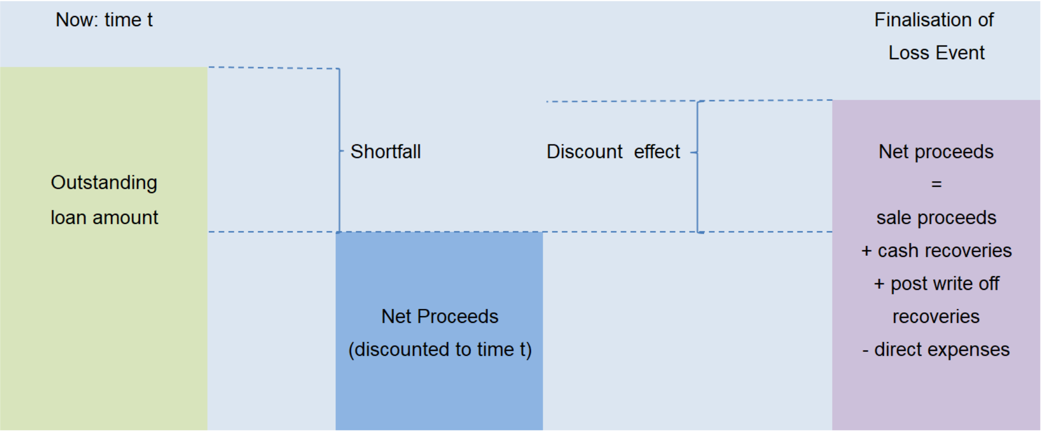
Now: time t
Finalisation of Loss Event
Outstanding loan amount
Shortfall
Discount effect
Net Proceeds (discounted to time t)
Net proceeds = sale proceeds + cash recoveries + post write off recoveries - direct expenses

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. It is assumed in [11] that h follows a normal distribution. This assumption is shown to be realistic in Section 4 below (see Figures 6 and 7 below). The following transformations are applied [11]

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

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

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

where $\Phi_z\,(D)$ and $\phi_z(D)$ represent the cumulative distribution function and probability density function of the standard normal distribution [11].

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.

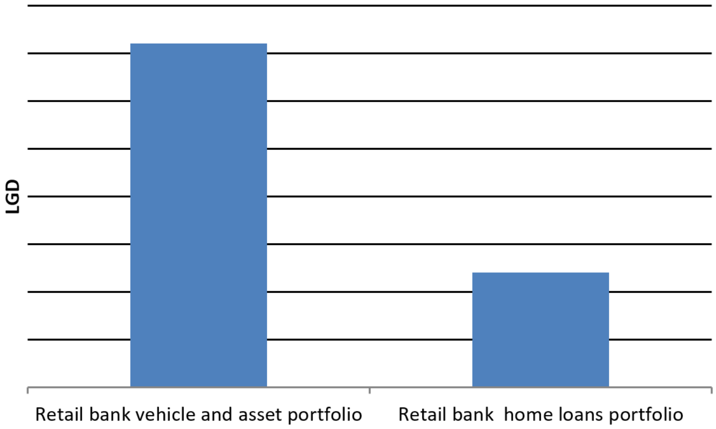

Figure 3: Retail bank data loss given default.

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.

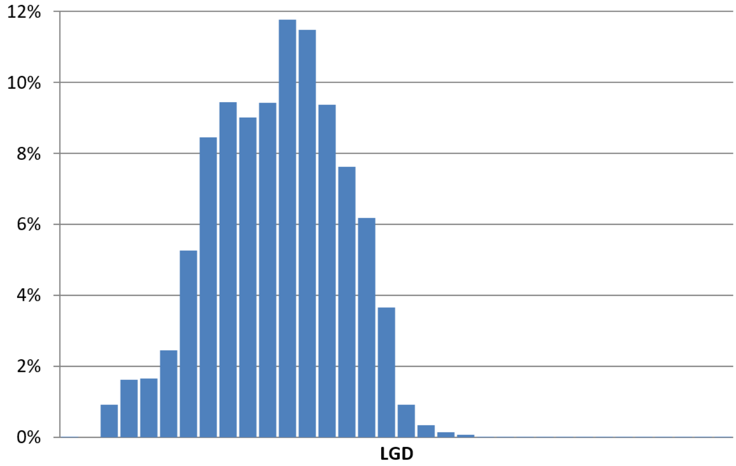

Figure 4: Vehicle and asset finance LGD distribution.

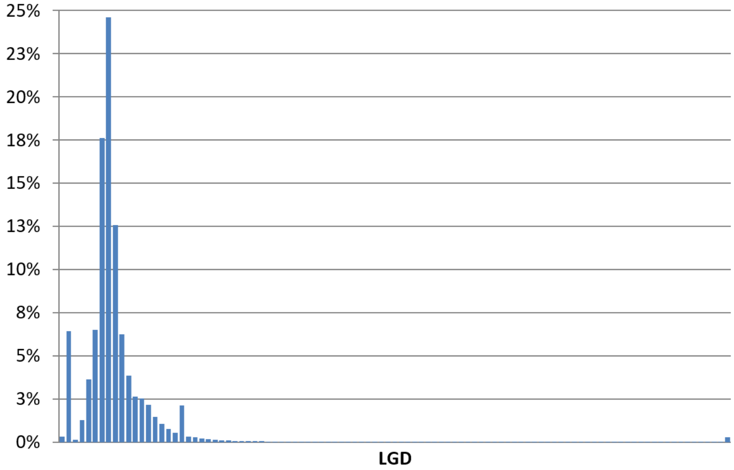

Figure 5: Home loans LGD distribution.

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 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.

	Test	Home Loans p value	Vehicle and asset finance 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.

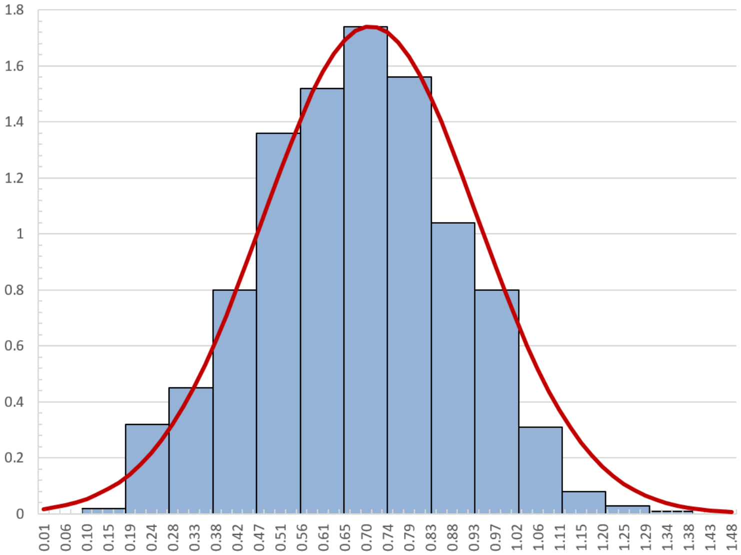

Figure 6: Vehicle and asset finance haircut distribution.

4.1.1 Censoring

Censoring is a peculiar feature of survival analysis and occurs when information is known over a certain interval [9]. 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 8 indicates the reference period data used for development and is used to describe censoring.

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

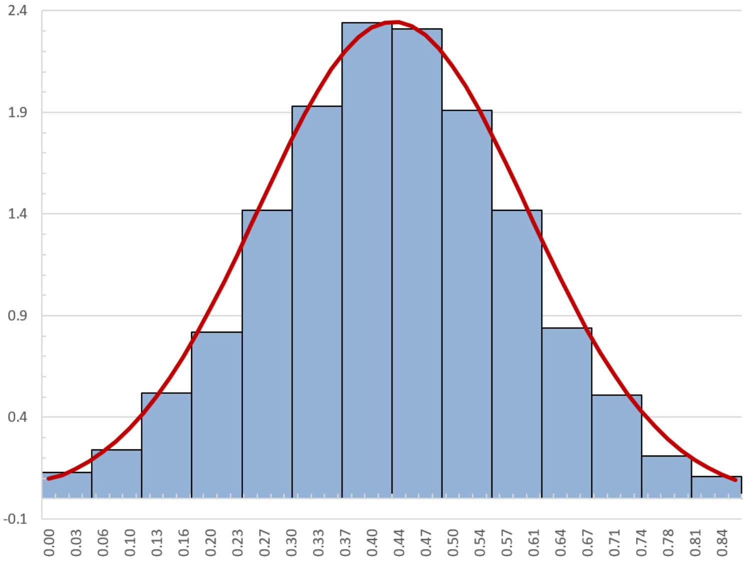

Figure 7: Home loans haircut distribution.

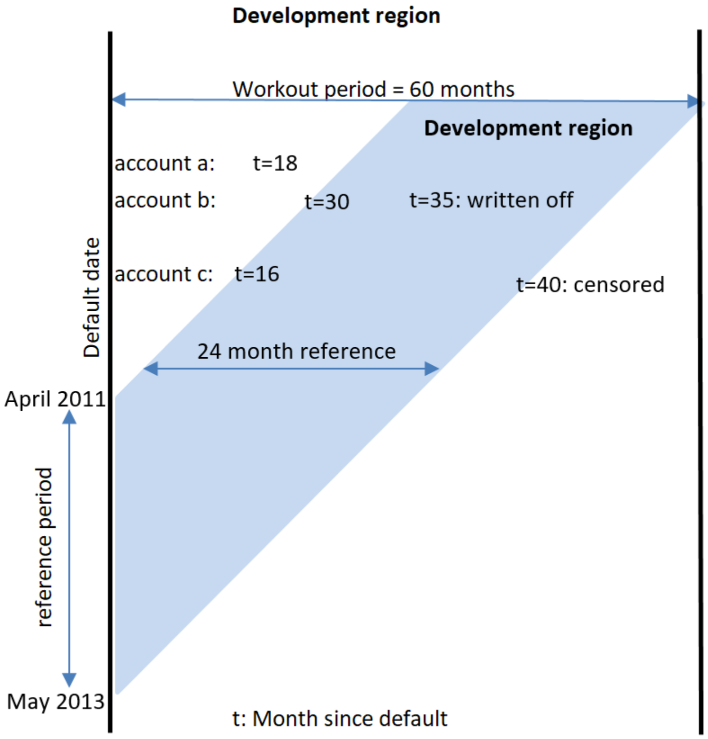

Figure 8: Development reference period data.

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 $t = 18$ [9].
- 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 $t = 30$ [9].
- 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 $t = 40$ [9].

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 $P_{i,t}\left(W_{\mathcal{T}}\right)$, $P_{i,t}\left(C_{\mathcal{T}}\right)$ and $P_{i,t}\left(I_{\mathcal{T}}\right)$ are estimated using the survival curves simulated 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, $S^{w}\left(t\right)$ and $S^{c}\left(t\right)$, were defined in Section 3.1 and combined to produce the three probabilities $P_{i,t}\left(W_{\mathcal{T}}\right)$, $P_{i,t}\left(C_{\mathcal{T}}\right)$ and $P_{i,t}\left(I_{\mathcal{T}}\right)$. 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 $S^{w}\left(t\right)$ and $S^{c}\left(t\right)$ are simulated for this paper. The article by [5] 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

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

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

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

where the cumulative hazard rate is taken as

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

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, $S\left(t\right)$, is written as

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

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

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

and it follows that

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

Next, let Y be a random variable with distribution function F. It follows that $U=F\left(Y\right)$ is uniformly distributed on the interval $\left[0,1\right]$, abbreviated as $U\ \sim\ Uni\left[0,1\right]$. 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

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

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

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

and U is a random variable with,

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

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

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

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

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

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,

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

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

and

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

The equation, $T\ \sim\ Exp(\lambda e^{\mathbf{x}'\beta})$, is used to simulate the survival time for both survival curves $S^w\left(t\right)$ and $S^c\left(t\right)$. 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 [5]. 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 [11]. 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. 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.

	Home loans		Vehicle and asset finance	
	$P_{i,t}(W_\tau)$	$P_{i,t}(C_\tau)$	$P_{i,t}(W_\tau)$	$P_{i,t}(C_\tau)$
λ	0.01	0.026	0.041	0.042
c	0.05	0.04	0.046	0.043

Table 2: Simulation study parameters.

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

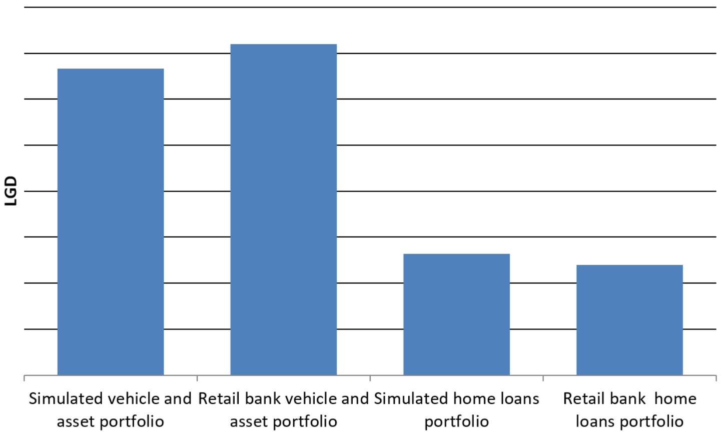

Figure 9: Simulated loss given default.

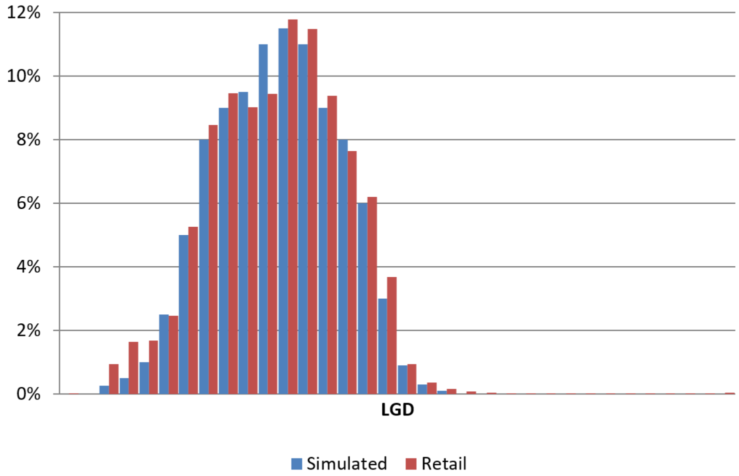

Figure 10: Vehicle and asset finance LGD distribution.

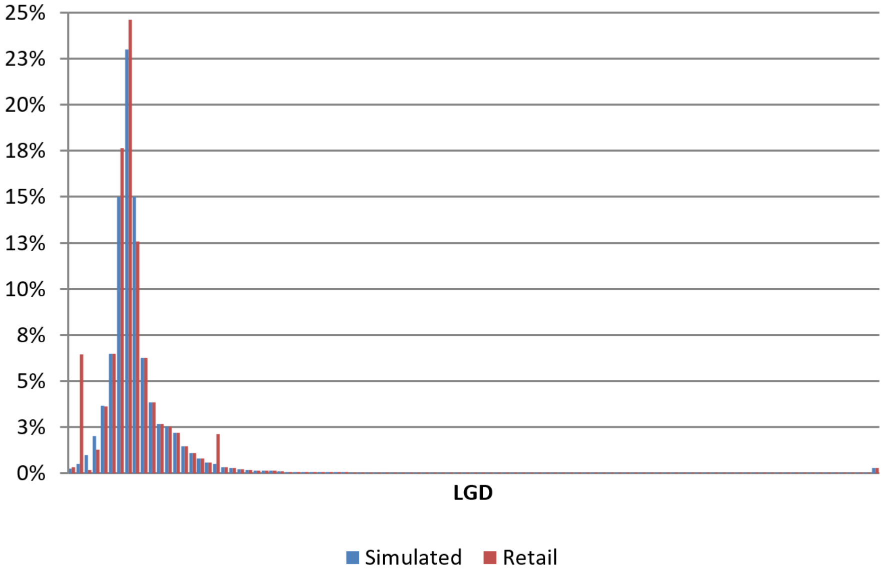

Figure 11: Home loans LGD distribution.

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.

5 Results

This section covers the results for both the retail and the simulated datasets. The method described by [11] to predict the probability components, $P_{i,t}(W_\tau)$, $P_{i,t}(C_\tau)$ and $P_{i,t}(I_\tau)$, 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.

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. 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.

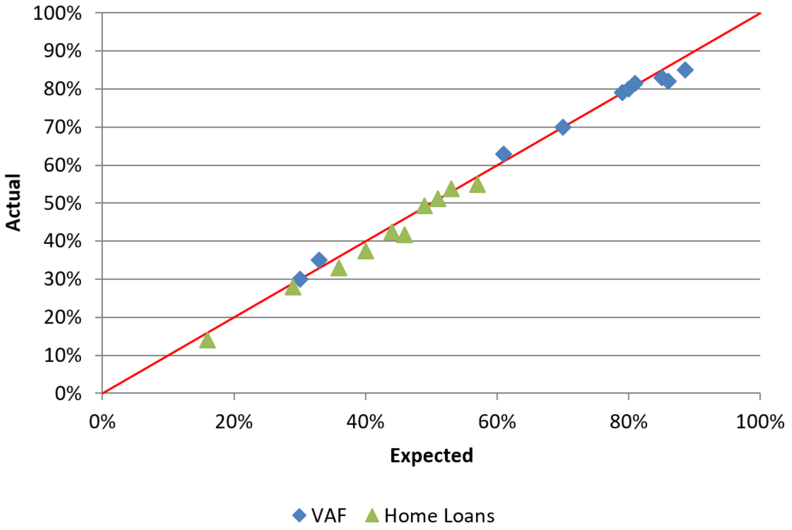

Figure 12: Haircut model accuracy on retail data.

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 re-

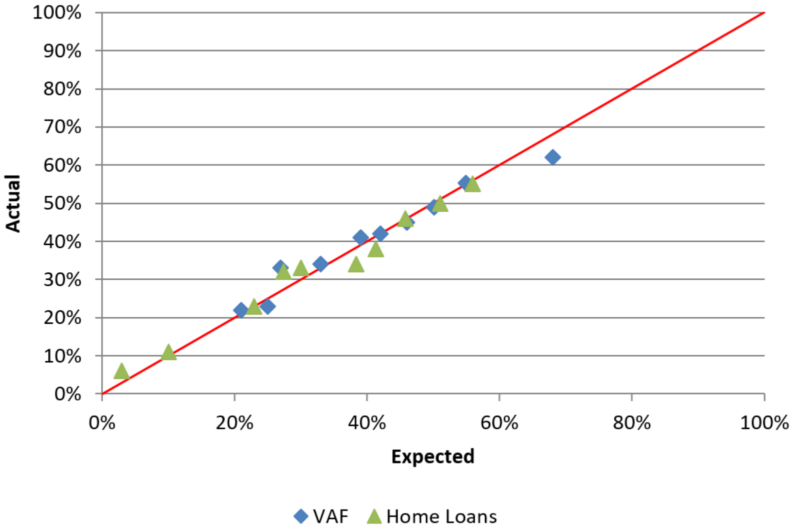

Figure 13: Probability of cure model accuracy on retail data.

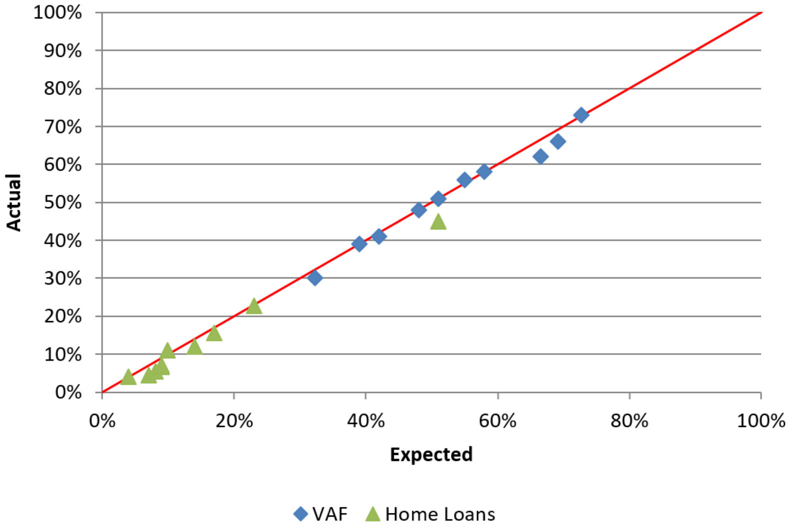

Figure 14: Probability of write-off model accuracy on retail data.

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

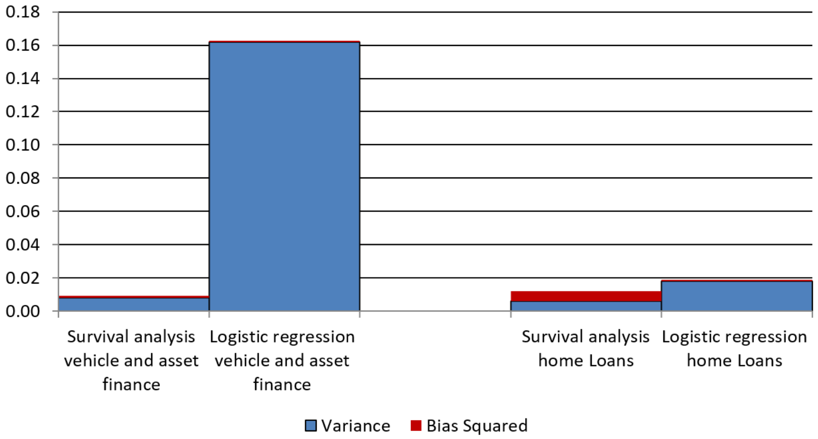

Figure 15: Retail data MSE, variance and bias.

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.

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.

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 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.

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.

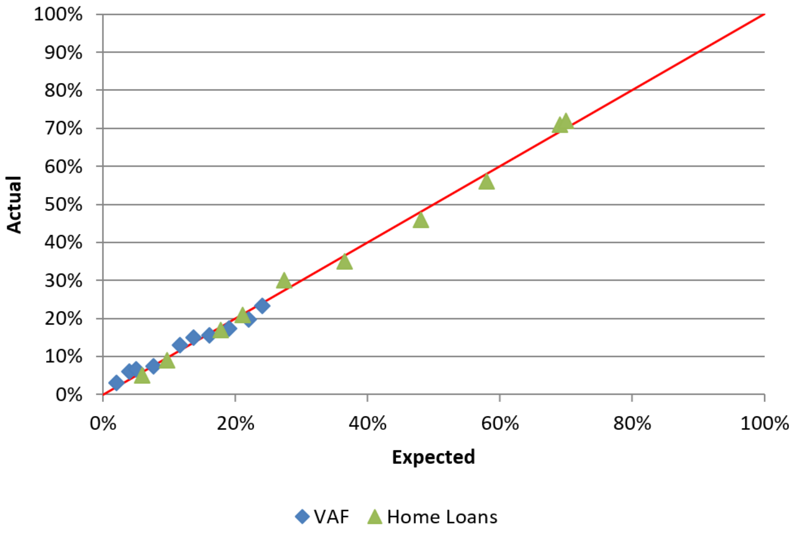

Figure 16: LGD accuracy on retail data.

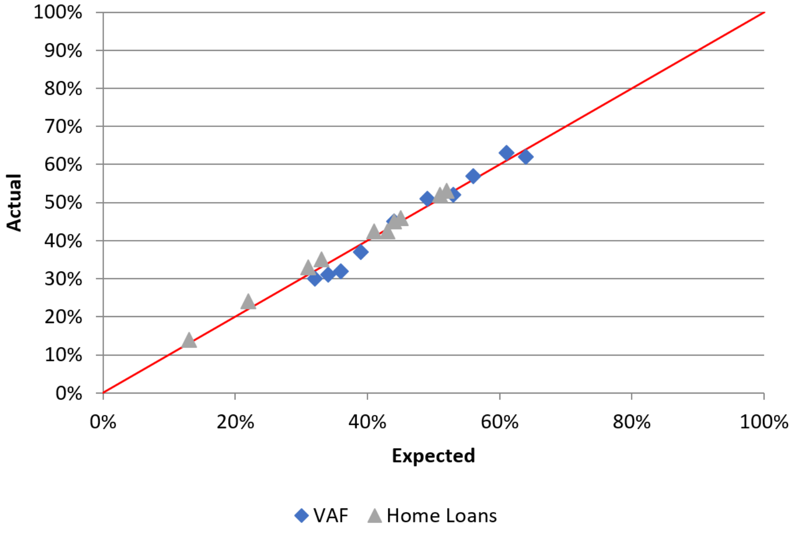

Figure 17: Haircut model accuracy on simulated data.

6 Conclusion

In the paper by [11] an indirect LGD modelling approach was described. The probability components were modelled by making use of logistic regression and the loss severity

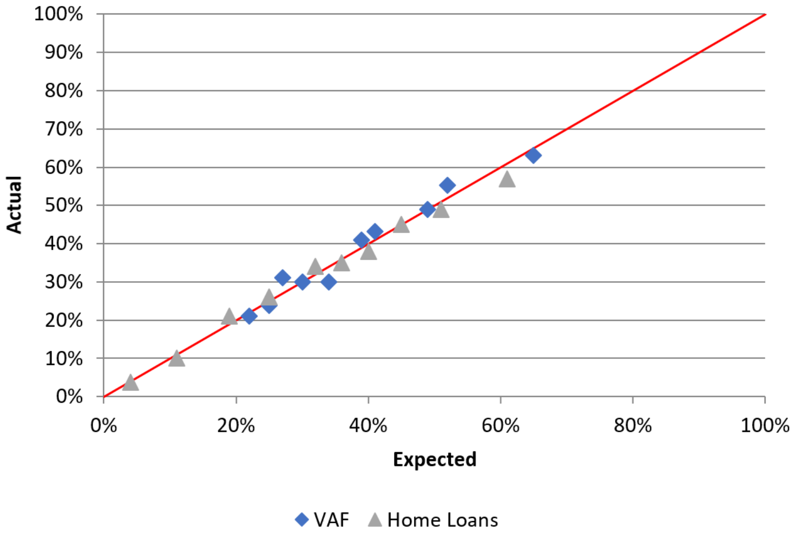

Figure 18: Probability of cure model accuracy on simulated data.

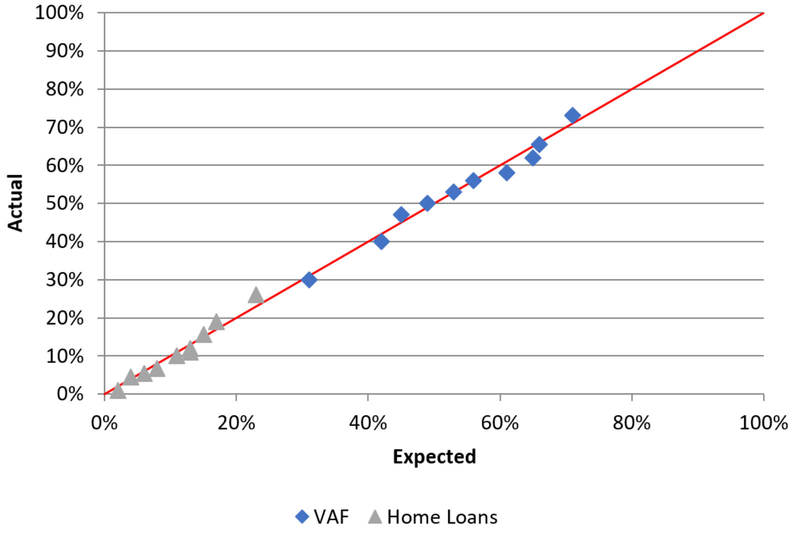

Figure 19: Probability of write-off model accuracy on simulated data.

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.

component was estimated by applying the haircut model.

This paper sets out to investigate whether predictability could be enhanced if the logistic regression for modelling the probability component was replaced by survival analysis.

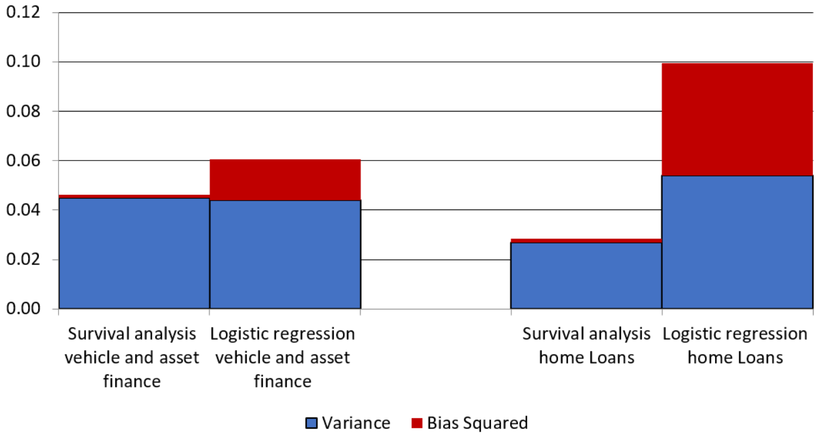

Figure 20: Simulated MSE, variance and bias.

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

Table 6: Simulated home loans portfolio.

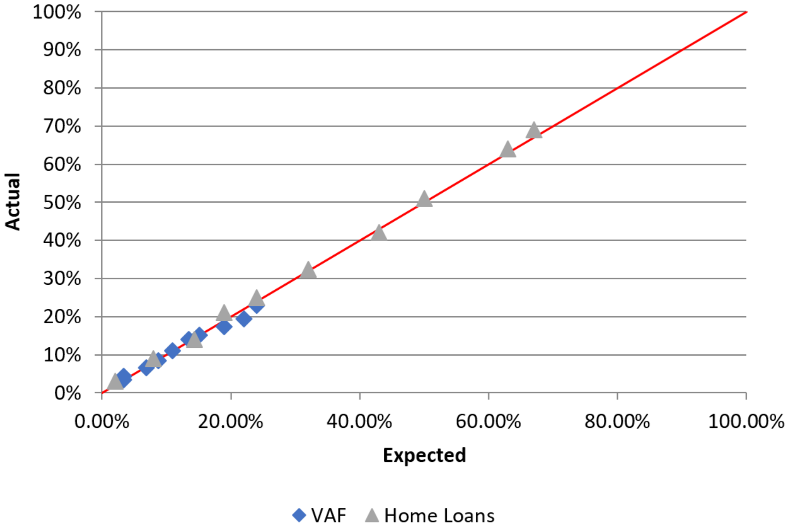

Figure 21: LGD accuracy on simulated data.

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.

In [11], the probability component was modelled 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 this 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 this 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 are the following:

- 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 [13] adapted Basel LGD modelling techniques to model the IFRS 9 LGD, future research could focus on extending this approach to IFRS 9 models.

References

[1] Anderson R, 2007, *The credit scoring toolkit*, Oxford University Press, Oxford.

[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(1)**, pp. 171–182.

[5] Bender R, Augustin T & Blettner M, 2005, *Generating Survival Times to Simulate Cox Proportional Hazards Model*, Statistics in Medicine, **31(29)**, pp. 3946–3958.

[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, **16(4)**, pp. 257–274.

[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, **28(1)**, 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(1)**, pp. 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)**, pp. 53–83.

[14] Qi M & Yang X, 2009, *Loss given default of high loan to value residential mortgages*, Journal of Banking and Finance, **33(5)** pp. 788–799.

[15] Schimdt K, 2006, *Methods and Models of Loss Reserving Based on Run-Off Triangles: A Unifying Survey*, Casualty Actuarial Society Forum, **(2)**, pp. 269–317

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

[17] Qi M & Zhao X, 2011, *Comparison of modelling methods for Loss Given Default*, Journal of Banking and Finance, **35(11)**, pp. 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(4)**, pp. 548–562.

[20] Witzany J, Rychnovsky M & Charamza P, 2012, *Survival Analysis in LGD Modelling*, European Financial and Accounting Journal, **7(1)**, pp. 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, **28(1)**, pp. 204–215.

[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.