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

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

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

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

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

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

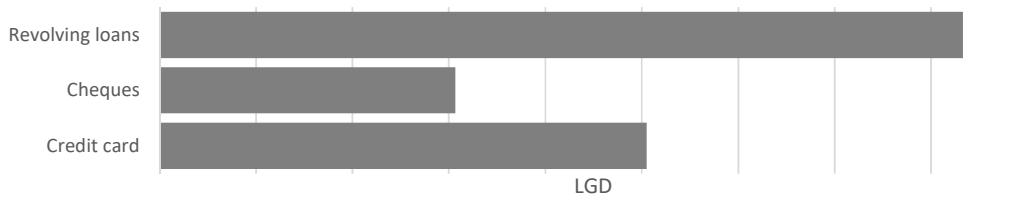

Figure 1. Retail banks datasets loss given default.

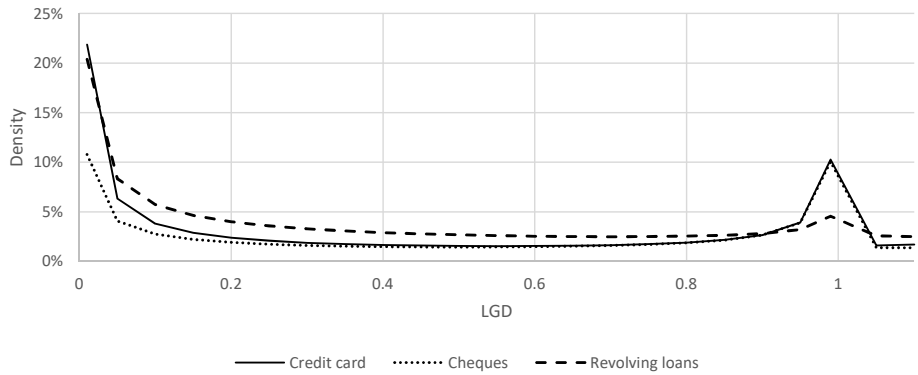

Figure 2. Retail banks datasets LGD distribution.

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.

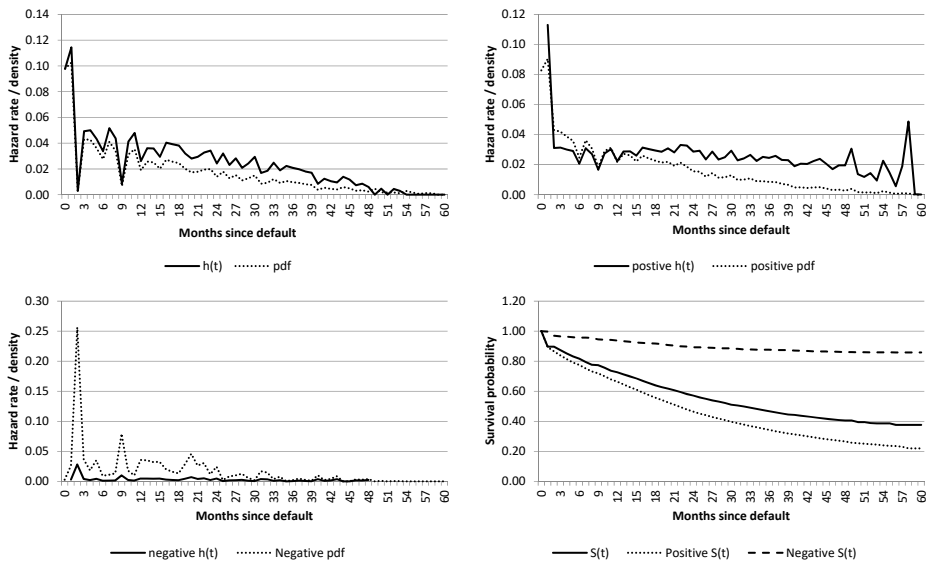

Figure 3. Credit card actual hazard rate, distribution and survival curve.

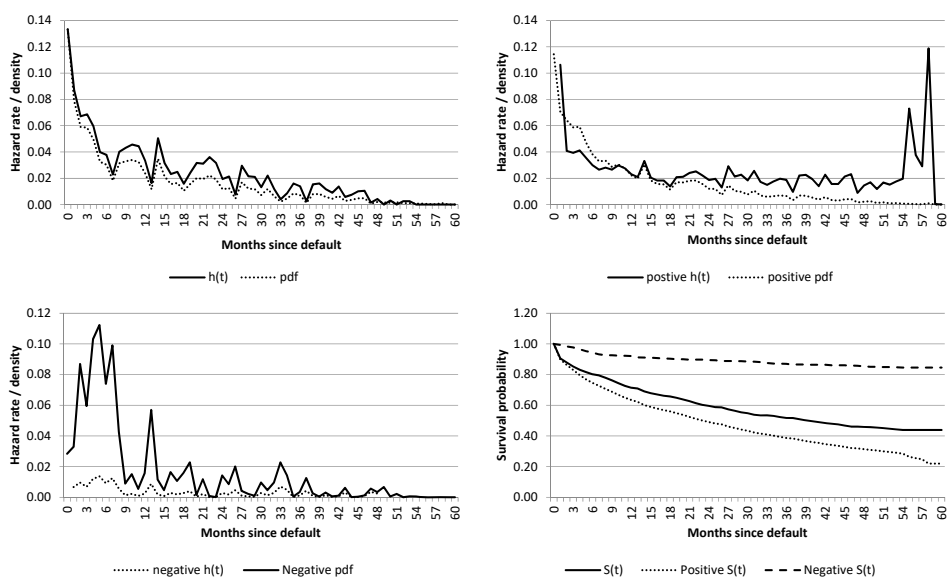

Figure 4. Cheque actual hazard rate, distribution and survival curve.

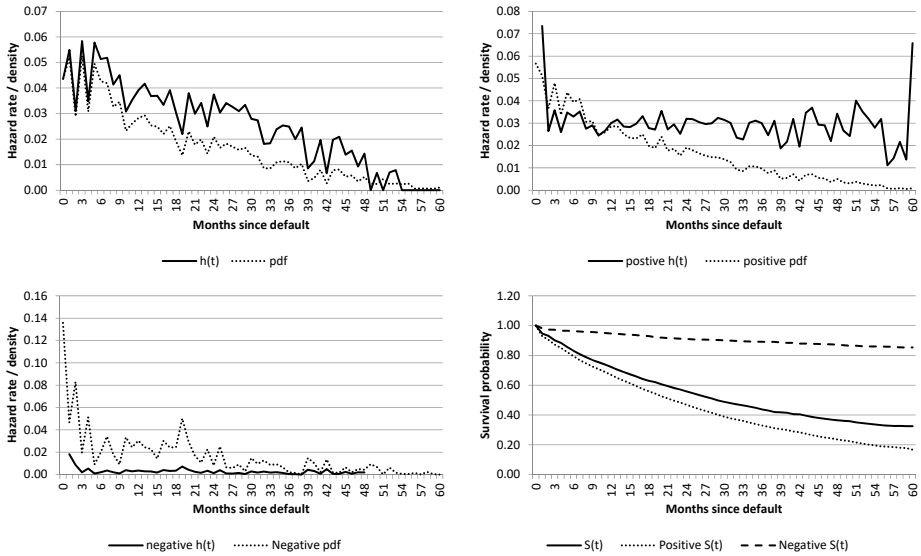

Figure 5. Revolving loan actual hazard rate, distribution and survival curve.

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.

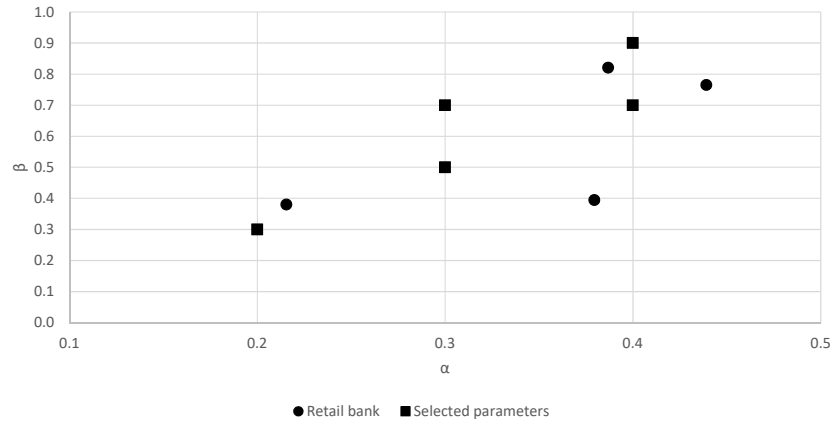

Figure 6. Beta distribution parameter estimates.

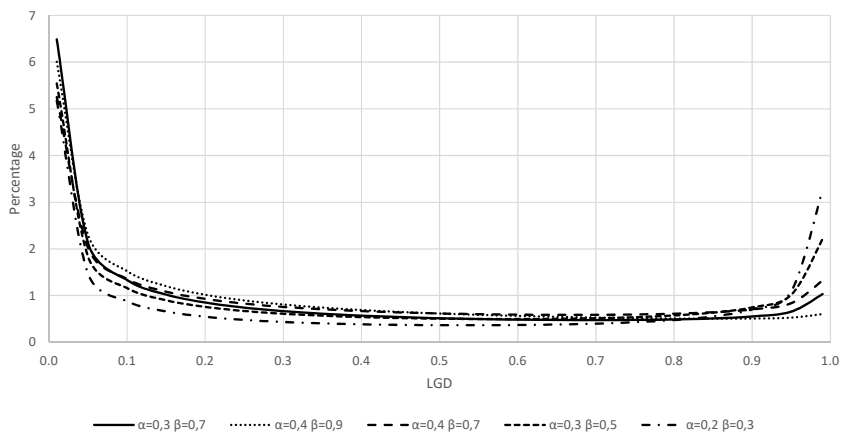

Figure 7. Beta distribution pdf.

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

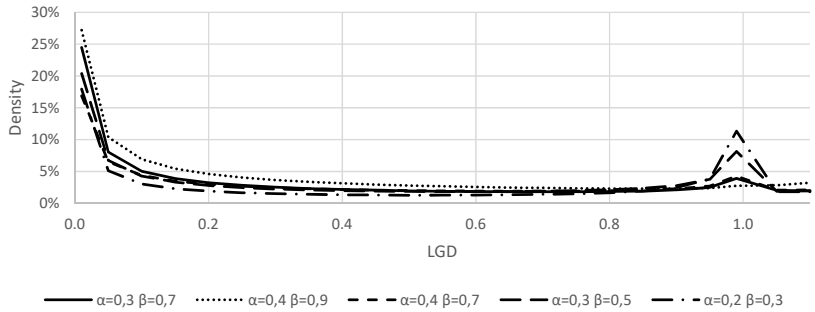

Figure 8. Beta distribution pdf with artificially added over-recoveries.

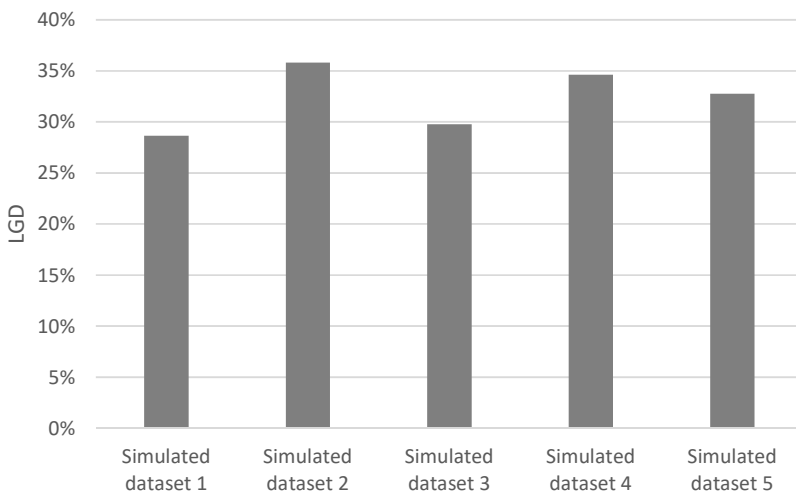

Figure 9. Simulated datasets loss given default.

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.

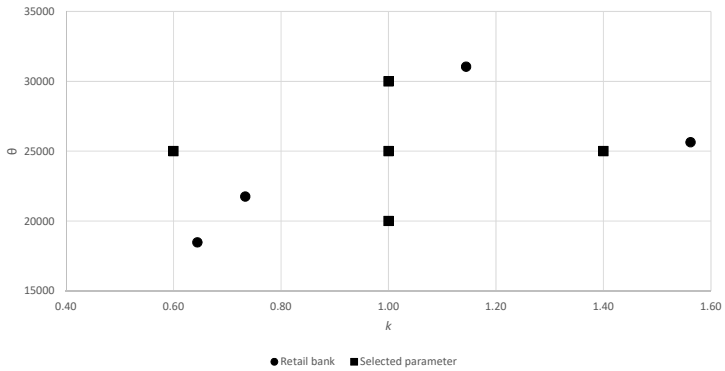

Figure 10. Gamma distribution parameter estimates.

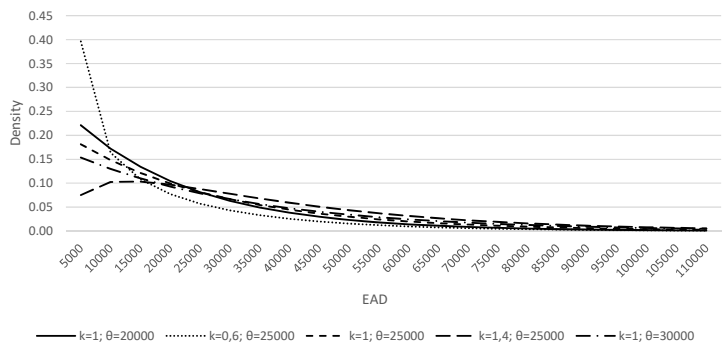

Figure 11. Gamma distribution pdf.

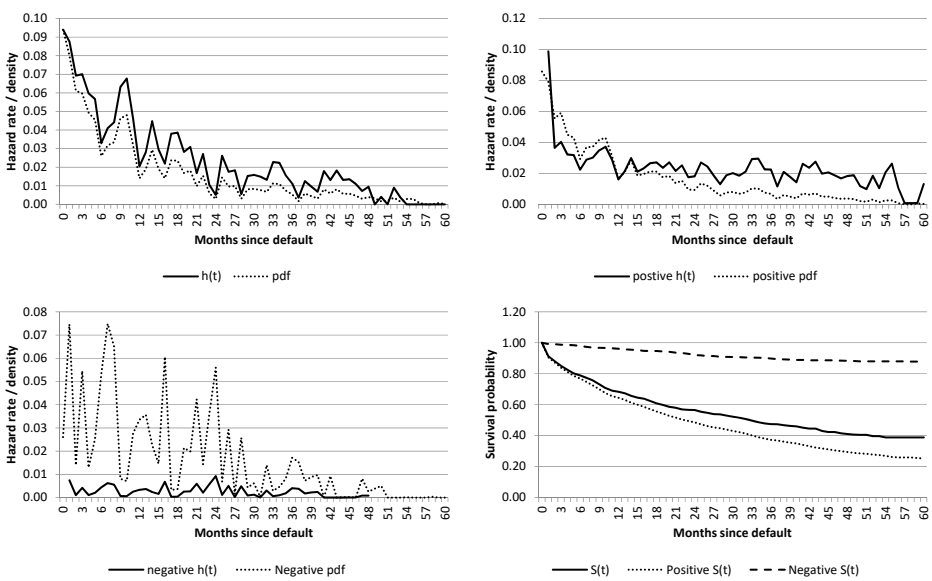

Figure 12. Simulated dataset 1 actual hazard rate, distribution and survival curve.

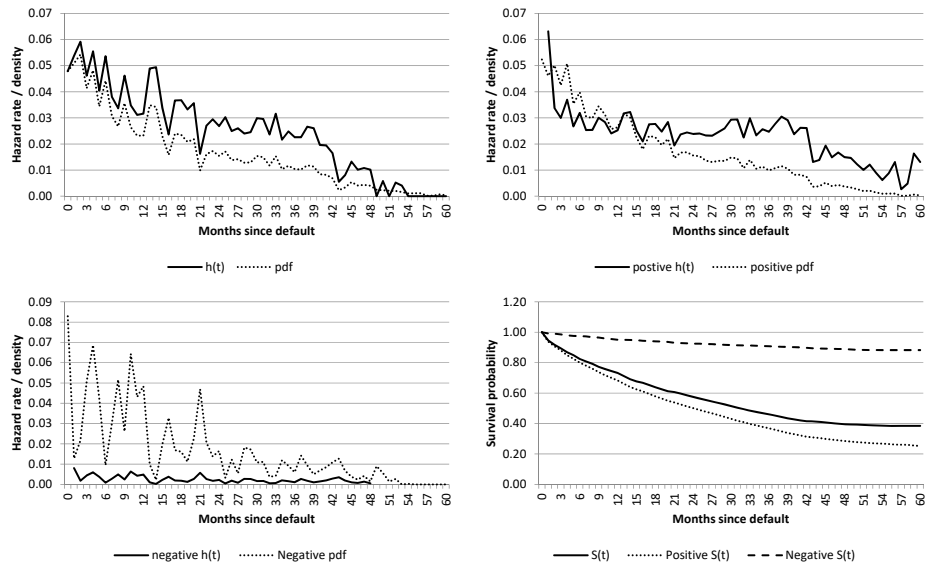

Figure 13. Simulated dataset 2 actual hazard rate, distribution and survival curve.

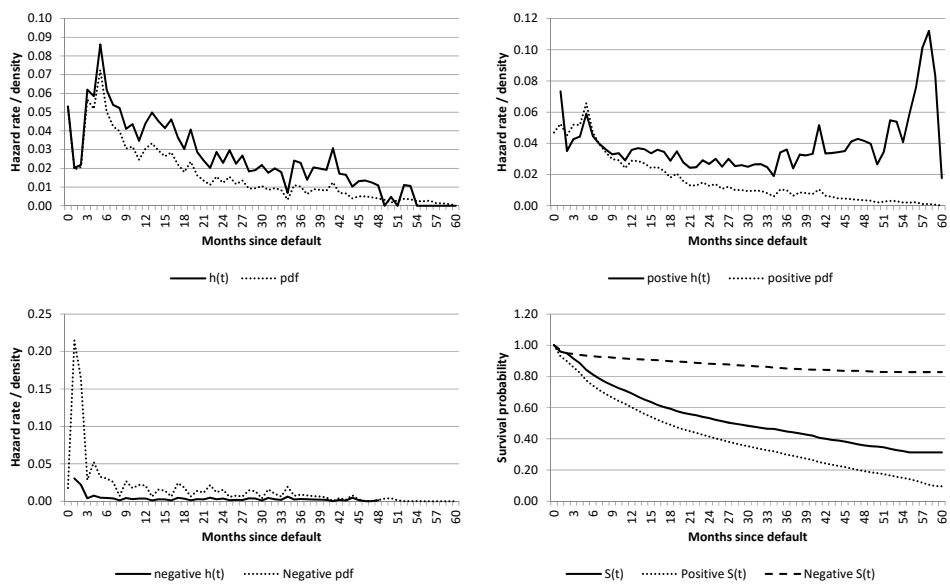

Figure 14. Simulated dataset 3 actual hazard rate, distribution and survival curve.

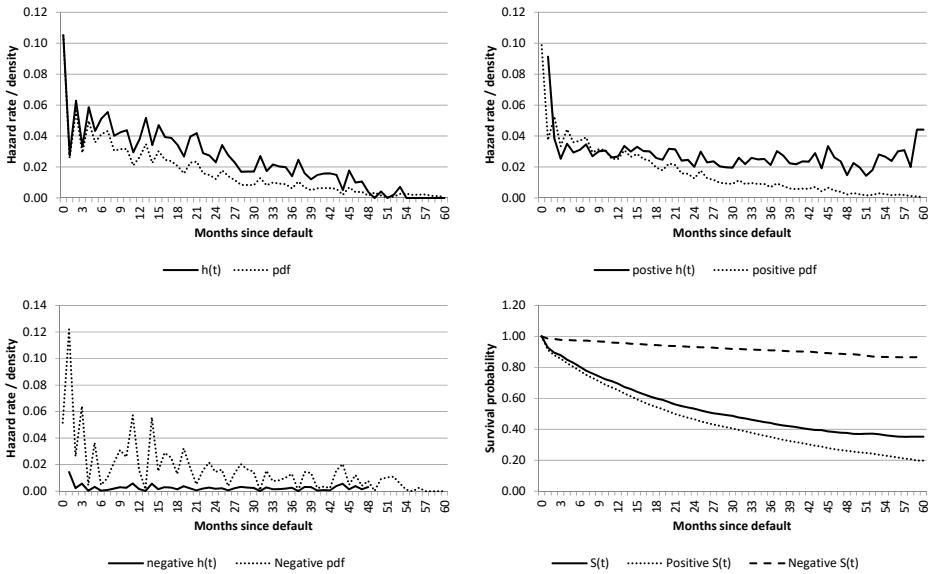

Figure 15. Simulated dataset 4 actual hazard rate, distribution and survival curve.

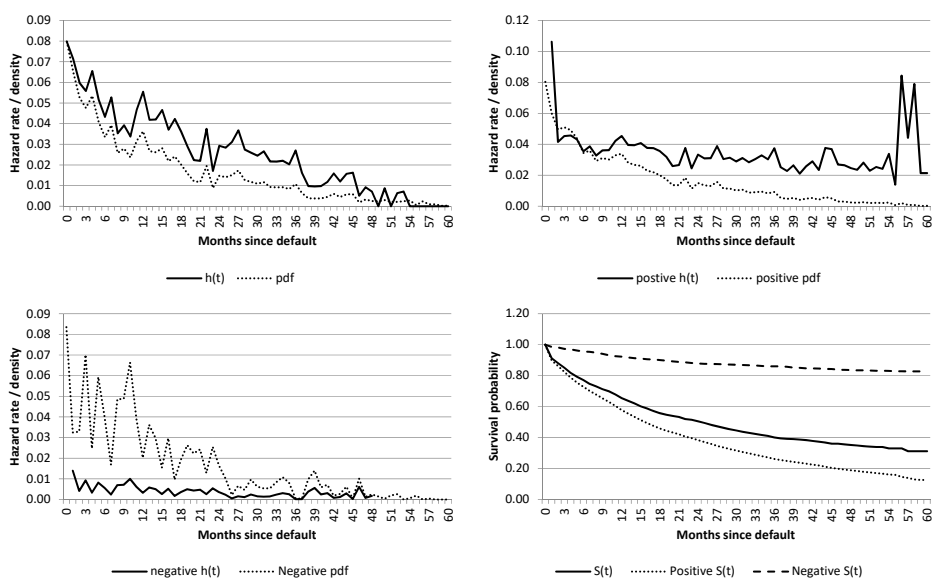

Figure 16. Simulated dataset 5 actual hazard rate, distribution and survival curve.

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:

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

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

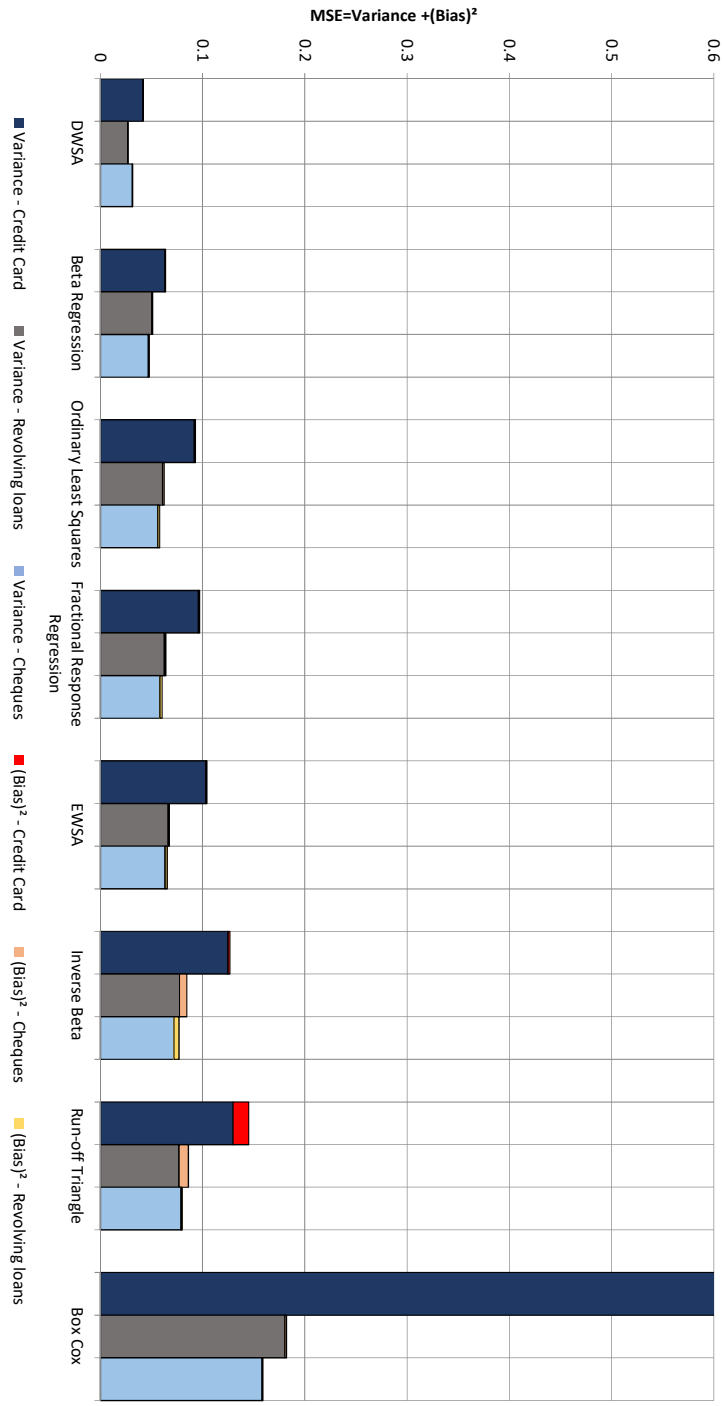

Figure 17. Retail bank dataset results for direct modelling approaches.

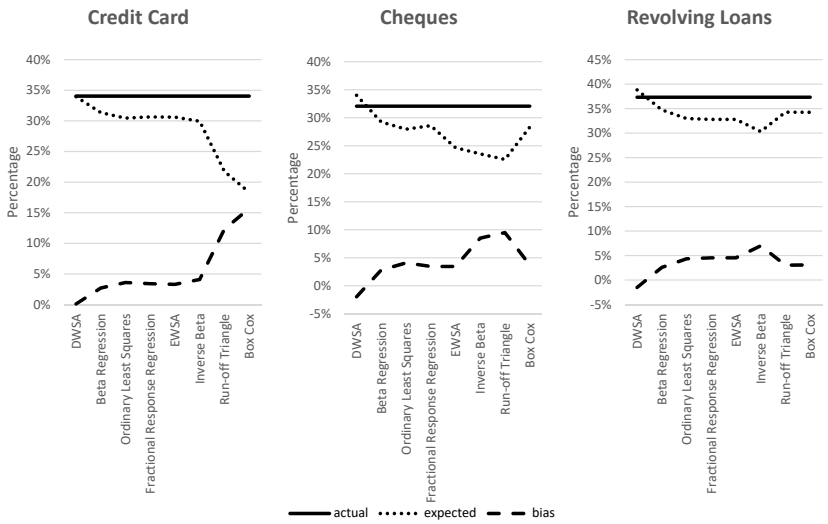

Figure 18. Bias, actual recovery rate and expected recovery rate for the retail bank datasets.

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

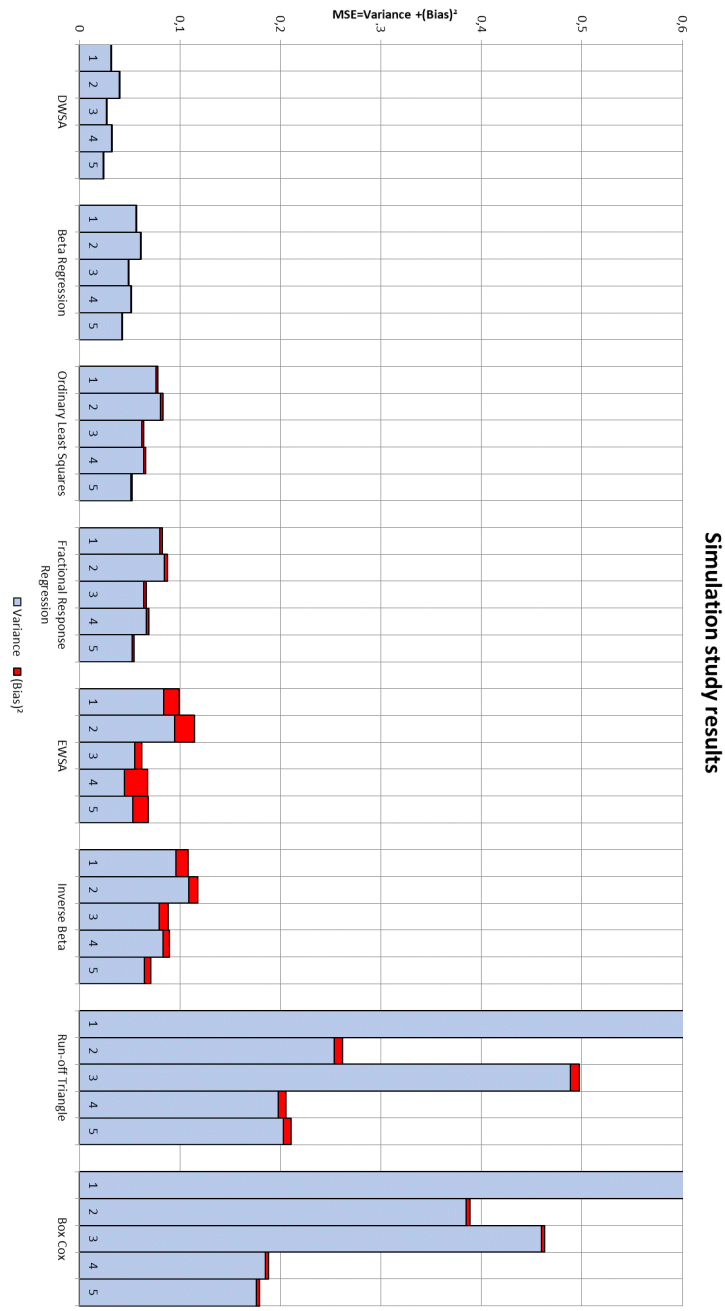

Figure 19. Simulation study results for direct modelling approaches.

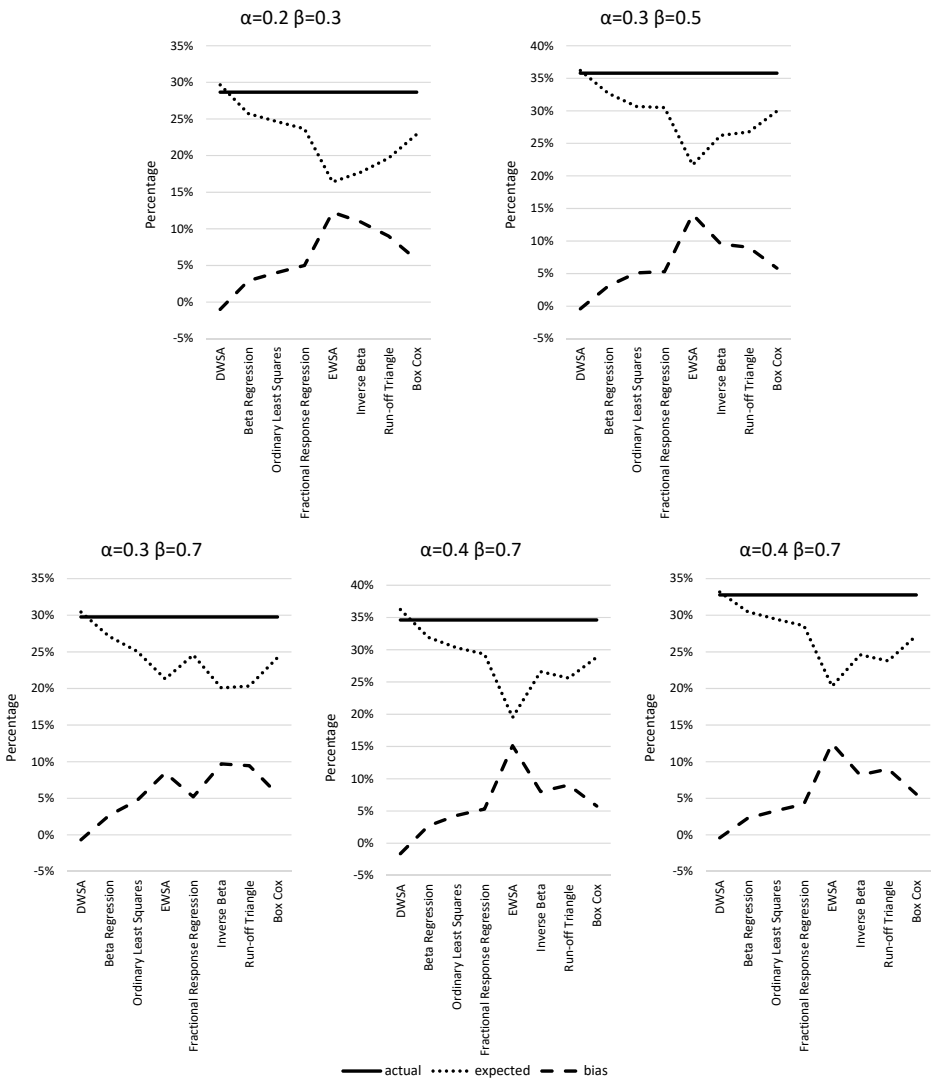

Figure 20. Bias, actual recovery rate and expected recovery rate for the simulated datasets.

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 $i = 2, ..., n$. 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).

	0	k	n
0	$C_{0,0}$	$C_{0,k}$	$C_{0,n}$
i	$C_{i,0}$	$C_{i,k}$	
n	$C_{n,0}$		

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.