BANK I KREDYT
BANK & CREDIT

Credit loss modelling using beta distribution in a Bayesian approach

Aneta Ptak-Chmielewska*, Paweł Kopciuszewski#

Submitted: 8 January 2024. Accepted: 7 March 2024.

Abstract

The Advanced Internal Rating Based (AIRB) approach is more and more frequently applied by banks. Bank analysts decide to use their own approach to calculate basic risk parameters such as Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD). The problem of small samples in LGD estimation is always a challenge for researchers and analytics. The paper proposes the basic LGD model based on splitting recoveries into two classes of recoveries: close to 0 or close to 1, and based on that split the construction of the LGD model with the combination of two binary models. The main advantage of the paper is, however, addressing the unresolved cases incorporated in the LGD estimation process by using a Bayesian approach which assumes a beta distribution of further recoveries for unresolved cases. An additional advantage of the paper is that the proposed modelling approach for LGD is illustrated on real data for mortgage loans for one of the European banks.

Keywords: Loss Given Default (LGD), Bayesian approach, beta regression, unresolved cases, small sample

JEL: C1, C11

* Warsaw School of Economics; e-mail: aptak@sgh.waw.pl.
# Vistula University of Warsaw, ING Hubs Poland.

1. Introduction and background

The most challenging problem in Loss Given Default estimation is the small sample (lack of data) and unresolved cases-observations (unfinished workout period). In specific portfolios such as mortgages or corporates, those challenges are always present and are not avoidable, and must be addressed.

In those situations, all statistical and econometric models based on large sample assumptions will not work properly. The estimates will always be biased due to small sample properties and will inadequately account for the censored, unresolved exposures in the portfolio.

According to the European Banking Authority EBA Guidelines (EBA/GL/2017/16) on estimation of risk parameters, such as PD or LGD, the defaulted exposure is given much attention BCBS (2017). LGD estimation should cover resolved cases as well as unresolved ones. In this context, the methodology proposed in this paper covers this urgent issue. As mentioned in the Guidelines, to have a realistic view of long-run average LGD unresolved (incomplete recovery) cases should be also included. The values of future recoveries are not observed, but should be estimated based on the already completed history of resolved cases. In this way, "long-run average LGD" will be more objective. It is also mentioned in the Guidelines that long-run LGD should be based on adjusted observed LGD also taking into account unresolved cases. This is the case when the moment of default is quite recent and the recovery process could not be finalized (i.e. it is shorter than the assumed maximum workout period). It is also important to mention that risk drivers such as length of recovery process and status of recovery process should also be included in LGD estimation as well as collateral. The proposed Bayesian approach addresses this challenge. This approach is focused on building one model including unresolved and resolved cases at the same level in the modelling process.

The modelling approach proposed in this paper is a two steps estimation procedure. The first step is based on splitting recoveries into two groups of recoveries: close to 0 or close to 1. This general observation of LGD distributions makes us construct an LGD model with the combination of two binary models. The second step addresses the challenge when building the LGD model – the use of unresolved cases in the estimation process. We apply a Bayesian approach which assumes a beta distribution of further recoveries for unresolved cases. The Bayesian approach is also considered with the LGD estimation for resolved cases with a proposed combination of two binary models.

The structure of the paper is as follows. Firstly, we provide a review of the existing literature about LGD estimation, as well as literature discussing the Bayesian approach for LGD estimation and small samples and unresolved cases as well. Secondly, we discuss the chosen modelling approach, while also providing the classical model estimation. The next section contains the results of the Bayesian model with beta distribution for LGD modelling. The final section provides the conclusions of our results and some suggestions for further research.

2. Literature review

In the literature there are many formulas and different approaches for LGD modelling in the literature. Many of them are based on Vasicek distribution or a similar approach. In Frye and Jacobs (2012) the LGD function connects the conditionally expected LGD rate (cLGD) to the conditionally expected default rate (cDR). In another proposal by Frye (2000) the recovery is a linear function of the normal risk factor associated to the Vasicek distribution. Pykhtin (2003) proposes parameterization of the amount, volatility, and systematic risk of a loan's collateral and infers the loan's LGD. Tasche (2004) assumes a connection between LGD and the systematic risk factor at the loan level. This idiosyncratic influence is integrated. Giese (2005) makes a direct specification of the functional form linking cLGD to cDR and Hillebrand (2006) introduces a second systematic factor that is integrated out to produce cLGD given cDR. In another proposal Giese (2006) uses the beta distribution on the systematic factor *Y*. Düllmann and Trapp (2004) propose the recovery rate to be modelled as a logit transformation of a normally distributed random variable *Y*.

The most popular approach in LGD modelling is just computing averages for selected homogenous groups-pools (Izzi, Oricchio, Vitale 2012), next is linear regression (Anolli, Beccalli, Giordani 2013; Loterman et al. 2012) or other types of regressions such as beta regression (Huang, Oosterlee 2011). This simple approach is of course preferred because it is clear and understandable for managers and authorities. Quite recently, new, more complicated and advanced approaches are proposed in the literature. Those proposals include machine learning methods, but also non-parametric methods as simple and more promising in results. Unfortunately, those "black-box" models are not preferred by regulatory bodies as they are not clear and understandable and interpretable for the customer. Recent innovative and interesting applications include decision trees as more interpretable (Belotti, Crook 2007), or neural networks (Brown 2012) and Markov chains (Luo, Shevchenko 2013), as well as scoring based methods (Van Berkel, Siddiqi 2012) or two-stage models (Brown 2012; Yao, Crook, Andreeva 2017; Papouskova, Hajek 2019) and the k-NN non-parametric approach (Ptak-Chmielewska et al. 2023).

Another approach for LGD estimation is just a market-based approach. For researchers the market data are relatively easily accessible. Implied LGD has been examined by some researchers using data from different countries and markets such as BBB rated US corporate bonds (Bakshi, Madan, Zhang 2001) or for Argentinian government bonds (Andritzky 2005). Results for market derived LGD for corporate bonds are characterized by high estimation errors and general low precision (Christensen, Henrik 2006). That is why an internal approach based on the workout period and empirically realized LGD is preferred. Unfortunately data for such estimates are not so easily accessible by academics and researchers.

The most difficult problem with LGD modelling is the issue of unresolved cases. Some results based on different portfolios show that it may take up to four years or longer from default to full recovery (Kosak, Poljsak 2010; Hurt, Felsovaly 1998). The research shows the potentially long time period before full recoveries are achieved. The simplest and therefore the most popular approach is just including incomplete cases in modelling and treating those cases as complete (Baesens, Roesch, Scheule 2016). Unfortunately this simplified approach may lead to a significantly different estimate of real LGD values. Ptak-Chmielewska, Kopciuszewski and Matuszyk (2023) proposed a non-parametric and survival approach for unresolved cases with promising results. The need to recognize the impact of unresolved cases is clear from research. Finally, the relationship between recovery rates and economic cycles has been researched by extrapolating recoveries for the LGD estimation for unresolved cases (Brumma, Urlich, Schmidt 2014). For any method, a discussion of the assumptions regarding the treatment of unresolved cases should also be provided according to Basel IV requirements (Nielsen, Roth 2017).

Data samples for LGD modelling in the small and medium enterprises (SME) segment or corporates are rather small data samples (due to small number of defaulted exposure) and the literature on that topic is also rare. There are some solutions discussed in the literature to manage the small sample issue.

Some of the proposals concern the use of external data or change the time period and incorporate future recoveries. One of the examples can be the proposal by Chalupka and Kopecsni (2009), who used methods based on realized losses to extend small samples. The authors applied their solution on the SME sector of the Czech Republic. They included a limited time period for recovery and assume a full recovery when the proportion of the exposure was sufficiently high. Another application was proposed by Zieba (2017) with an overview of methods of how to increase the sample size. This work was based on a real LGD data sample and the author proved that the proposed extrapolation of recovery rates is the most efficient way to increase sample size and improve the precision of LGD estimation.

The quite limited examples of the Bayesian approach for LGD modelling also support the value added and new insight into this research area proposed by this paper. One of the potential examples could be LGD estimation for the most difficult portfolios such as unsecured loans. A typical approach in such a situation is just an estimation of two different models and a final calculation based on a two--step approach. However, this can be problematic because results must be combined together to make the final predictions about LGD. The advantage of using the Bayesian approach in this situation is that only a single hierarchical model can be estimated instead of two separate models. One of the examples of using the Bayesian approach was found in the work by Bijak and Thomas (2015). The authors used Bayesian methods and the frequentist approach for comparison. They applied their approach to the data on personal loans from one of the largest UK banks. The posterior estimates of means of parameters that have been calculated by the authors using the Bayesian approach were very similar to the values calculated in the frequentist approach. The basic advantage of using the Bayesian model in this case was the individual predictive distribution of LGD for each loan. The authors also calculated the so-called down-turn LGD and stressed LGD.

The biggest value added of the proposed approach in this paper is the utilization of the Bayesian approach. We propose the theoretical and methodological elaboration and application on real banking data. For the Bayesian approach we elaborated and used beta distribution.

3. Basic logistic LGD models

The basic assumption is that overall LGD for the both resolved and unresolved cases is a random variable. Assuming a distribution for the observed LGD allows to incorporate additional knowledge to the model which is contained in the distribution parameters. As total recoveries for resolved cases are known but are partially known for the unresolved ones, the final formula has to be different. In the classical approach the resolved cases are used to derive the model formula. Next, the unresolved ones are used to adjust the developed model. The current purpose in the paper is to present this new approach and test it with real data. The direct inclusion of the entire population in the modelling process makes the final model less biased. The currently proposed formula is the first attempt to apply it to LGD modelling and check the results and the possibility of using it to solve other LGD issues. The structure of the model allows much more information to be taken into account, such as external knowledge of the independent variables, relations between coefficients, distributions for all parameters included in the model, and the extension of the model structure towards any needs.

The data used for the modelling process comes from the years 2008 to 2018. They included 1,867 observations, of which 314 are unresolved, and 400 explanatory variables. The data was cleaned before being used for the modelling purpose in terms of excluding unintuitive variables, removing outliers, and replacing missing information.

The main inputs to Bayesian modelling are two logistic models which predict LGD = 0 and LGD = 1. These two models were built on the abovementioned data limited to resolved cases only. An important phase of modelling was the removal of strongly correlated variables and those with various multivariate and univariate relationships with the target variable. The models contain many application and behavioural variables at the customer and collateral level. Detailed descriptions about the estimation of logistic models were elaborated in a previous paper (Ptak-Chmielewska, Kopciuszewski 2023) and for convenience also included briefly in the appendix.

Bayesian modelling allows data for resolved and unresolved cases to be combined, but all the information from the data is provided by these two logistic models.

4. Beta Bayesian model for the total recoveries

Let us introduce the following basic markings, but notice that when referring to the end of the recovery process, this also means an observation of no more than the maximum workout period.

EAD_i is the exposure at Default for the *i*-th customer, it is assumed to be a customer related deterministic value.

R_{obs} is the observed recovery reduced by costs and other charges discounted at the Default moment, it is assumed to be a random variable.

R_{total} is the total recovery reduced by costs and other charges discounted at the Default moment, it is assumed to be a random variable, notice that it is equal to the R_{obs} for the resolved cases.

R_{add} is the observed additional future recovery unobserved for unresolved cases, reduced by costs and other charges discounted at the Default moment, it is assumed to be a random variable. It is assumed to be 0 for resolved cases.

RR_{obs} is the observed recovery rate up to the end of the observation period, i.e. the end of the recovery process for resolved cases and by the censored time for unresolved cases.

RR_{total} is the overall recovery rate up to the end of the recovery process for the both resolved cases and unresolved cases.

Suppose further that the total recovery R_{total} can be predicted from the linear model based on the two logistic models developed in the previous section as follows:

$$R_{total} = a_0 + a_1 \cdot LGD_0 + a_2 \cdot LGD_1 + \varepsilon_{total} \tag{1}$$

where $\varepsilon \sim \mathrm{F}(0,\ \sigma_{total}^2)$, F is any distribution family and σ_{total}^2 is the variance of the total recovery.

Then the expected mean $E(R_{total})$ equals the following linear formula with three degrees of freedom:

$$E(R_{total}) = a_0 + a_1 \cdot LGD_0 + a_2 \cdot LGD_1 \tag{2}$$

Suppose that R_{add} can be predicted from the linear model based on the basic information on the time (t) spent in the default state, the inclusion of which is recommended by the regulations, and additionally other explanatory variables:

$$R_{add} = (T - t) \cdot f\left(\sum_{j=1}^{k} b_j \cdot x_j\right) + \varepsilon_{add} \tag{3}$$

where:

$\varepsilon_{add} \sim \mathrm{F}(0,\ \sigma^2_{add}),\ F$ is any distribution family,

σ^2_{add} is the variance of the additional unobserved recovery,

t is the time spent in default up till the censored time moment,

T is the maximum workout period (90 months) for unresolved cases and the duration between the last month in the recovery process and default date for resolved cases, hence $t = T$ for resolved cases and consequently R_{add} is 0,

$x_1, ..., x_k$ are the explanatory variables predicting the future part of recoveries,

$b_1, ..., b_k$ are the parameters at explanatory variables in this formula,

$f(\cdot)$ is the function of the linear combination of the k explanatory variables.

Then the three random variables $R_{obs},\ R_{total}$ and R_{add} can be combined into one formula, assuming that the variance of the R_{obs} is $\sigma^2_{obs} = \sigma^2_{total} + \sigma^2_{add}$ and ε_{add} and ε_{total} are independent.

$$R_{obs} = R_{total} - R_{add} \tag{4}$$

Take into account that R_{obs} is the observed recovery in the experiment for both the resolved and unresolved cases, while R_{total} is the total recovery amount and hence it is equal to R_{obs} for resolved cases. R_{add} equals 0 for resolved cases and is the precited part of recovery for unresolved ones. All these three variables are treated as random. The main idea of the model is to combine all resolved and unresolved cases into one mathematical formula even if the two groups of cases are not consistent.

Let's introduce now the recovery ratio for the observed part of the recoveries as follows:

$RR_{obs,i} = \dfrac{R_{obs,i}}{EAD_i}$ for i-th facility is assumed to be a variable with the beta distribution

$$RR_{obs,i} \sim B(\alpha_i, \beta_i),\ \alpha_i > 0,\ \beta_i > 0 \tag{5}$$

Let's make the mean and the variance of the variable dependent on the above parameters α_i, β_i. Note that all calculations are for the whole portfolio, including both the resolved and unresolved cases. It is calculated conditionally under the given two binary models and additional explanatory variables including the time spent in default up till the censored time moment.

$$E\left(RR_{obs,i} \mid LGD_{0,i}, LGD_{1,i}, t_i, x_{1,i}, \ldots, x_{k,i}\right) = \frac{E(R_{total,i} - R_{add,i} \mid LGD_{0,i}, LGD_{1,i}, t_i, x_{1,i}, \ldots, x_{k,i})}{EAD_i} \tag{6}$$

$$E\left(RR_{obs,i} \mid LGD_{0,i}, LGD_{1,i}, t_i, x_{1,i}, \ldots, x_{k,i}\right) = \frac{\alpha_i}{\alpha_i + \beta_i} \tag{7}$$

Formula (6) results from the definition of the variable $RR_{obs,i}$, while formula (7) results from the beta density function assumed for the variable.

The beta family is the most appropriate family of distributions for the LGD parameter given its range of values and the wide class of distributions in that family.

Combining formulas (2), (3), (5) and (6), we get:

$$\frac{\alpha_i}{\alpha_i+\beta_i}=\frac{a_0+a_1\cdot LGD_{0,i}+a_2\cdot LGD_{1,i}-\left(T-t_i\right)\cdot f\left(\sum_{j=1}^{k}b_j\cdot x_{j,i}\right)}{EAD_i} \tag{8}$$

From the definition of R_{add} it follows that $t_i = T$ for the resolved cases and it is 0 in the above formula.

Similarly, another equation can be derived from the variance formula for the beta distribution:

$$Var\left(RR_{obs,i}\mid LGD_{0,i},LGD_{1,i},t_i,x_{1,i},\ldots,x_{k,i}\right)=\frac{Var\left(R_{total,i}-R_{add,i}\right)}{\left(EAD_i\right)^2}=\frac{\sigma_{obs}^2}{\left(EAD_i\right)^2} \tag{9}$$

$$Var\left(RR_{obs,i}\mid LGD_{0,i},LGD_{1,i},t_i,x_{1,i},\ldots,x_{k,i}\right)=\frac{\alpha_i\cdot\beta_i}{\left(\alpha_i+\beta_i\right)^2\cdot\left(\alpha_i+\beta_i+1\right)} \tag{10}$$

Formula (9) results from the definition of the variable $R_{obs,\ i}$, while formula (10) results from the beta density function assumed for the variable.

And next, combining formulas (9) and (10), we get:

$$\frac{\alpha_i\cdot\beta_i}{\left(\alpha_i+\beta_i\right)^2\cdot\left(\alpha_i+\beta_i+1\right)}=\frac{\sigma_{obs}^2}{\left(EAD_i\right)^2} \tag{11}$$

Given both equations, parameters α_i and β_i can be determined by data values and regression parameters $\beta_1,\ldots,\beta_k$. The current goal is not to derive the exact formulas for them because these equations can be incorporated into one Bayesian model without explicitly specifying α_i and β_i.

However α_i and β_i can be determined from formulas (7) and (10) as follows:

$$\alpha_i=\left(\frac{E_i\left(1-E_i\right)}{V_i}-1\right)E_i,\quad \beta_i=\left(\frac{E_i\left(1-E_i\right)}{V_i}-1\right)\left(1-E_i\right) \tag{12}$$

where E_i, V_i are the expected value and the variance of recovery rate $RR_{obs,\ i}$ conditionally under $LGD_{0,i}, LGD_{1,i}, t_i, x_{1,i},\ldots,x_{k,i}$. The E_i and V_i in turn can be given by the following formulas:

$$E_i=\frac{a_0+a_1\cdot LGD_{0,i}+a_2\cdot LGD_{1,i}-\left(T-t_i\right)\cdot f\left(\sum_{j=1}^{k}b_j\cdot x_{j,i}\right)}{EAD_i},\quad V_i=\frac{\sigma_{obs}^2}{\left(EAD_i\right)^2} \tag{13}$$

Taking into account formulas (12) and (13), beta parameters can be determined both using linear model parameters (formulas (1) and (3)) and explanatory variables from data.

5. Results of the Beta Bayesian model

The set of potential variables for LGD prediction was determined on the basis of other calculations, including correlation coefficients and temporary simpler models, and then the following variables were selected as potential ones for the final LGD model predicting the future recoveries:

EAD – Exposure at Default,
LTV_dynamics – Loan to Value dynamics: change compared to the previous period,
dpd_12 – days past due in last 12 months,
MTH_SINCE_LIMIT_START – months since start of using limit,
LTV_current – Loan to Value current,
past_due_amt_avg – average amount overdue,
F_BALANCE_RATIO – financial balance ratio in percent.

The SAS algorithm to discover the final LGD formula assuming the best choice from the set of explanatory variables is as follows:

All prior distributions were assumed to be noninformative. Two of them for parameters a2 and b0 were assumed to be positive as follows:

```
%let pars = a0 a1 a2 b0 b1,
%let r_add = b0 + b1 * F_BALANCE_RATIO,
ods graphics on,
proc mcmc data = data_lgd_bayes_beta ntu = 1000 nmc = 20000 propcov = quanew
diag = ALL outpost = post_kalib seed = 10 plot=density dic monitor = (&pars),
ods select PostSumInt mcse ess TADpanel densityPanel,
parms &pars 1,
R_add = max(0, & r_add),
E = (a0 + a1 * lgd_0 + a2 * lgd_1-n_months_to_def_end * R_add)/ead;
V = (&s * &s)/(ead * ead) (* &s is the empirical variance *)
alpha = max(0.0001, (E*(1-E)/V-1)*E);
beta = max(0.0001, (E*(1-E)/V-1)*(1-E));
prior a0 ~ uniform(-100, 100);
prior a1 ~ uniform(0, 100);
prior a2 ~ uniform(-100, 0);
prior b0 ~ uniform(0, 100);
prior b1 ~ uniform(-100, 100);
model rr_obs ~ beta(alpha, beta);
run.
```

Table 1 shows the final model posterior parameters and the quality of the model measured by R2. The discriminatory power of the model could only be checked for the resolved cases. The best quality is achieved for the model with f_balance_ratio as the explanatory variable. R2 is around 24% and it proves the good model quality in comparison with the results in the LGD literature. In addition, looking at the posterior distribution of the model parameters, it seems that f_balance_ratio is the best choice as well because the variance for a_0, a_1, a_2 parameters is not large and the distribution is quite smooth, symmetric and not concentrated on the border of the interval.

The model with EAD as the explanatory variable was unstable and the procedure yielded no results. Other models with more than one variable also gave poor results.

Table 1

Posterior parameters and the quality of the selected models

Variable	Parameter	Mean	Standard deviation	95% HPD interval		R2
None	a_0	0.5321	0.0112	0.5077	0.5517	15.39%
	a_1	0.5076	0.0121	0.4840	0.5309	
	a_2	-0.1514	0.0321	-0.1878	-0.0757	
	b_0	0.00149	0.000469	0.000683	0.00235	
LTV_dynamics	a_0	0.5497	0.00815	0.5342	0.5655	18.88%
	a_1	0.1115	0.0246	0.0537	0.1543	
	a_2	-0.2388	0.0113	-0.2613	-0.2171	
	b_0	0.0114	0.00331	0.00667	0.0174	
	b_1	-0.0165	0.00459	-0.0247	-0.0100	
Dpd_12	a_0	-0.0557	0.00221	-0.0602	-0.0514	< 0
	a_1	0.00261	0.00249	3.965E-7	0.00769	
	a_2	-0.00041	0.000407	-0.00124	-1.61E-7	
	b_0	16.1514	12.0823	0.0282	39.1365	
	b_1	-53.0651	21.9138	-90.8928	-12.4377	
MTH_SINCE_LIMIT_ START	a_0	2.8917	0.1322	2.7137	3.1646	< 0
	a_1	0.0281	0.0418	0.000067	0.0983	
	a_2	-3.3368	0.2534	-3.8506	-2.9759	
	b_0	0.2116	0.1135	0.0897	0.4569	
	b_1	-99.3539	0.8823	-99.9959	-98.0733	
LTV_current	a_0	2.2932	0.1003	2.1543	2.4776	< 0
	a_1	0.000557	0.000565	4.978E-7	0.00166	
	a_2	-2.4464	0.0628	-2.5326	-2.3301	
	b_0	0.0781	0.0100	0.0577	0.0901	
	b_1	-0.1291	0.0339	-0.1775	-0.0683	
past_due_amt_avg	a_0	0.5496	0.00807	0.5334	0.5647	18.87%
	a_1	0.1201	0.0236	0.0693	0.1640	
	a_2	-0.2345	0.0119	-0.2571	-0.2114	
	b_0	0.000366	0.000321	6.485E-8	0.00101	
	b_1	-46.4239	29.5188	-93.9533	-0.00367	
F_BALANCE_RATIO	a_0	0.5495	0.00800	0.5334	0.5650	24.25%
	a_1	0.4893	0.00955	0.4719	0.5091	
	a_2	-0.2364	0.0115	-0.2583	-0.2134	
	b_0	7.2793	3.9723	0.0637	14.3804	
	b_1	-56.6705	25.5214	-99.8236	-14.7716	

Figures 1–7 present the posterior densities for the model parameters. These figures and Table 1 constitute the basis for selecting the most appropriate explanatory variable predicting the future recovery.

Figure 1
Intercept

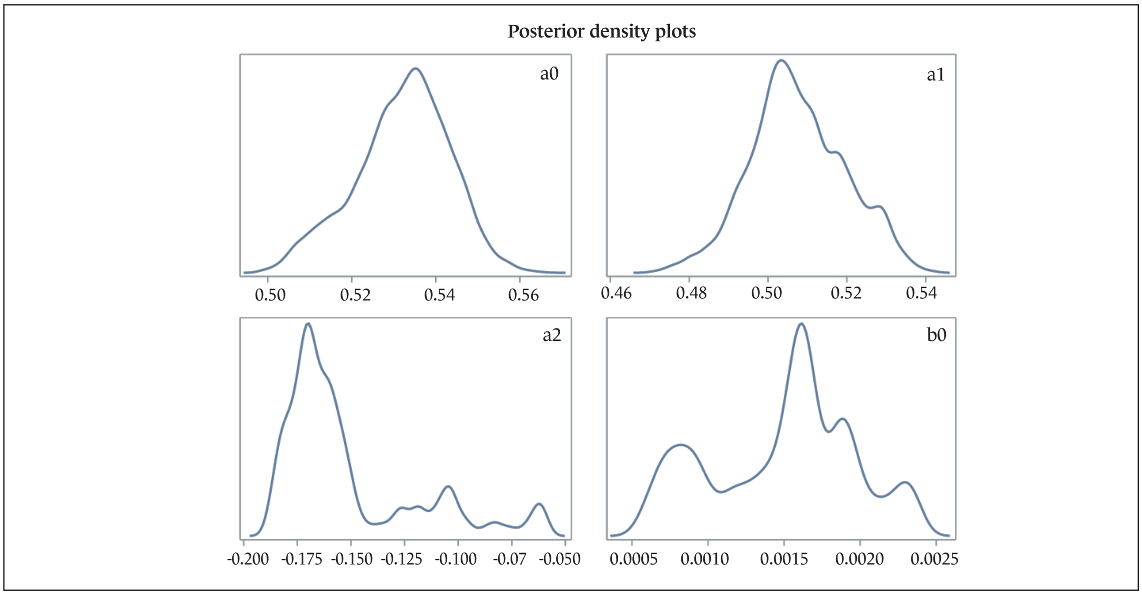


Figure 2
LTV_dynamics

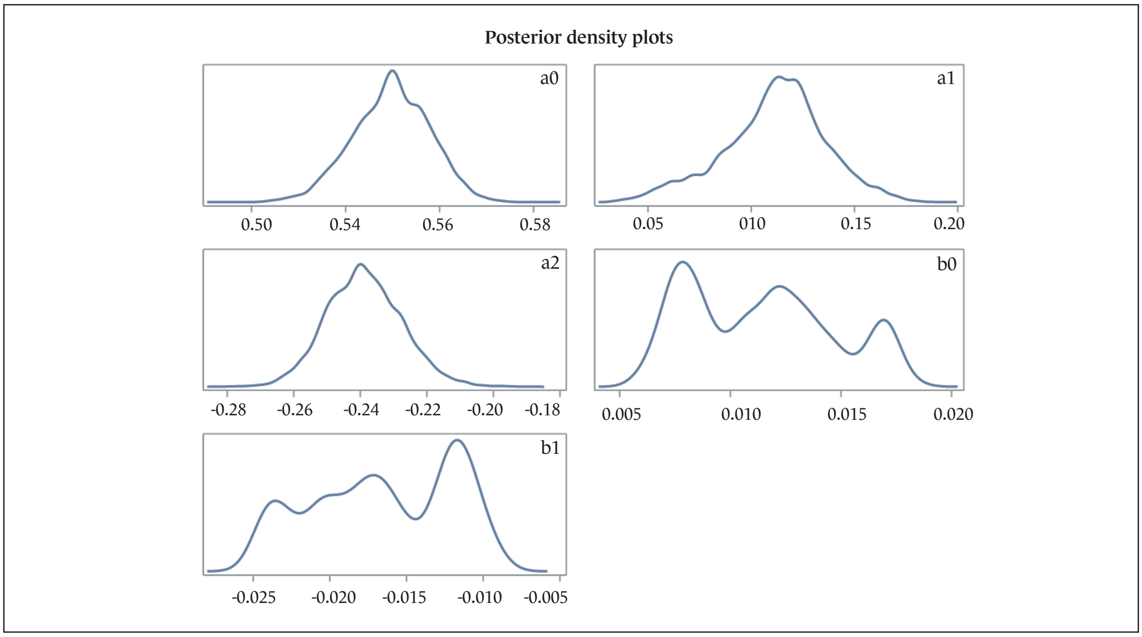


Figure 3
dpd_12

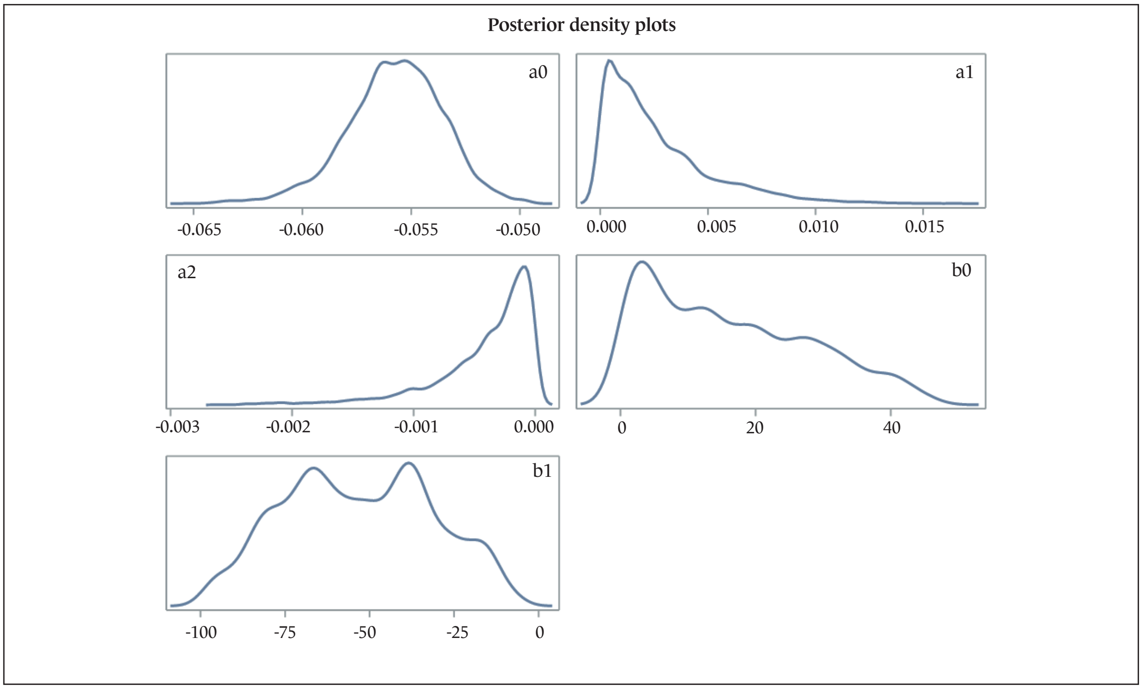


Figure 4
mth_since_limit_start

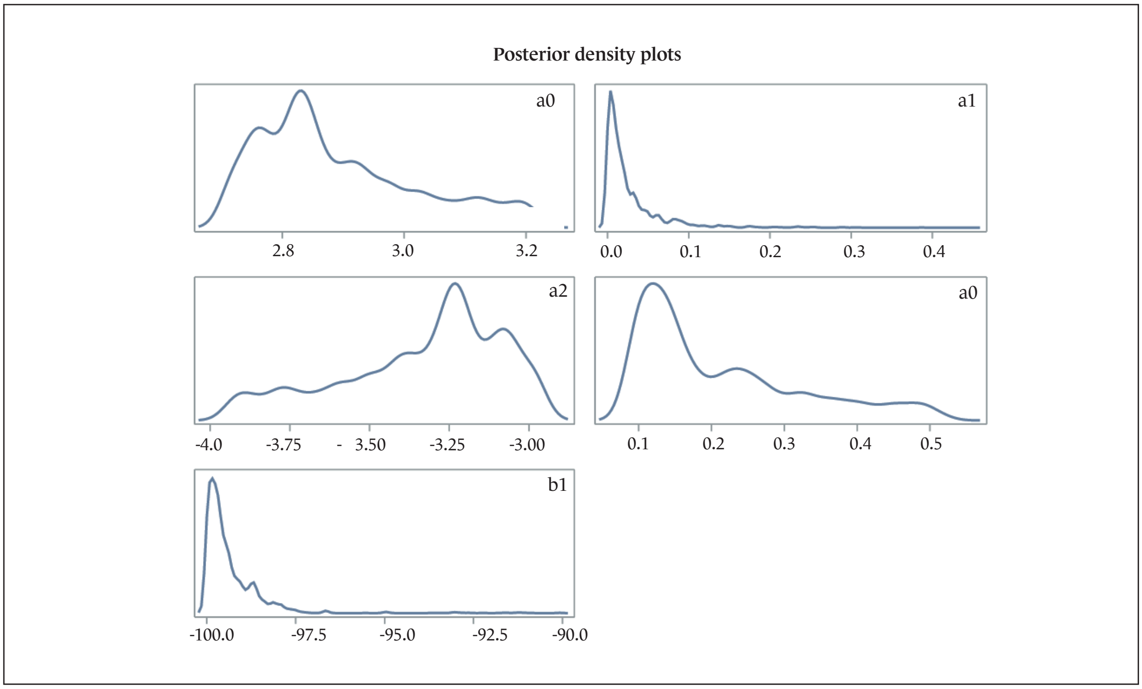


Figure 5
LTV_current

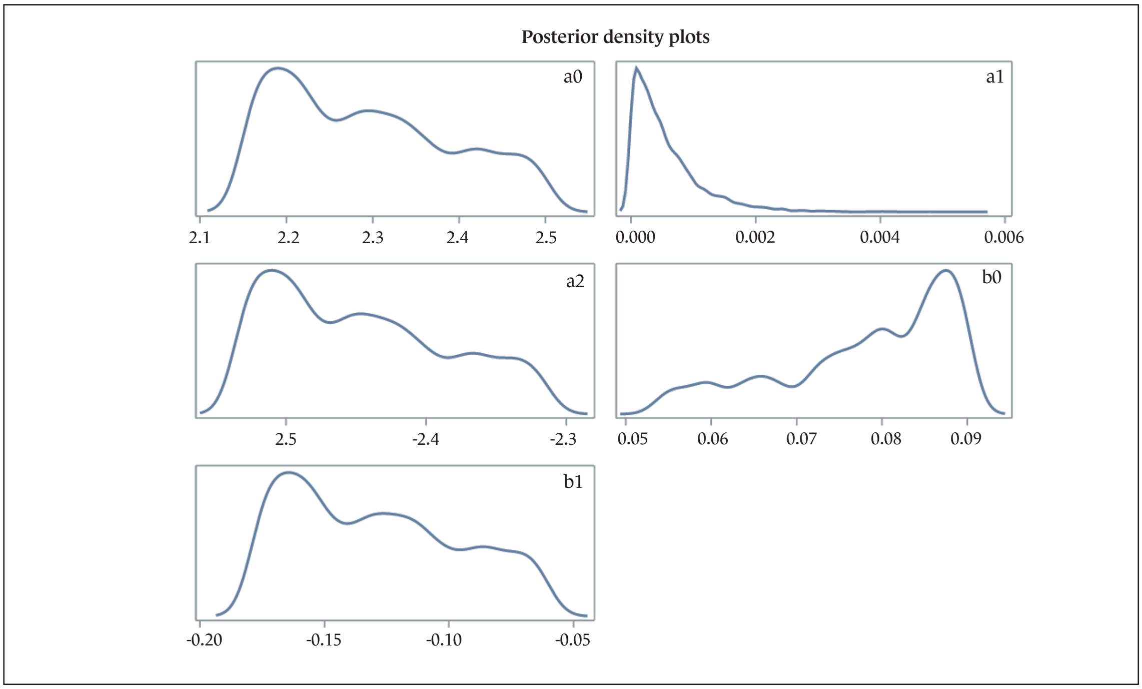


Figure 6
past_due_amt_avg

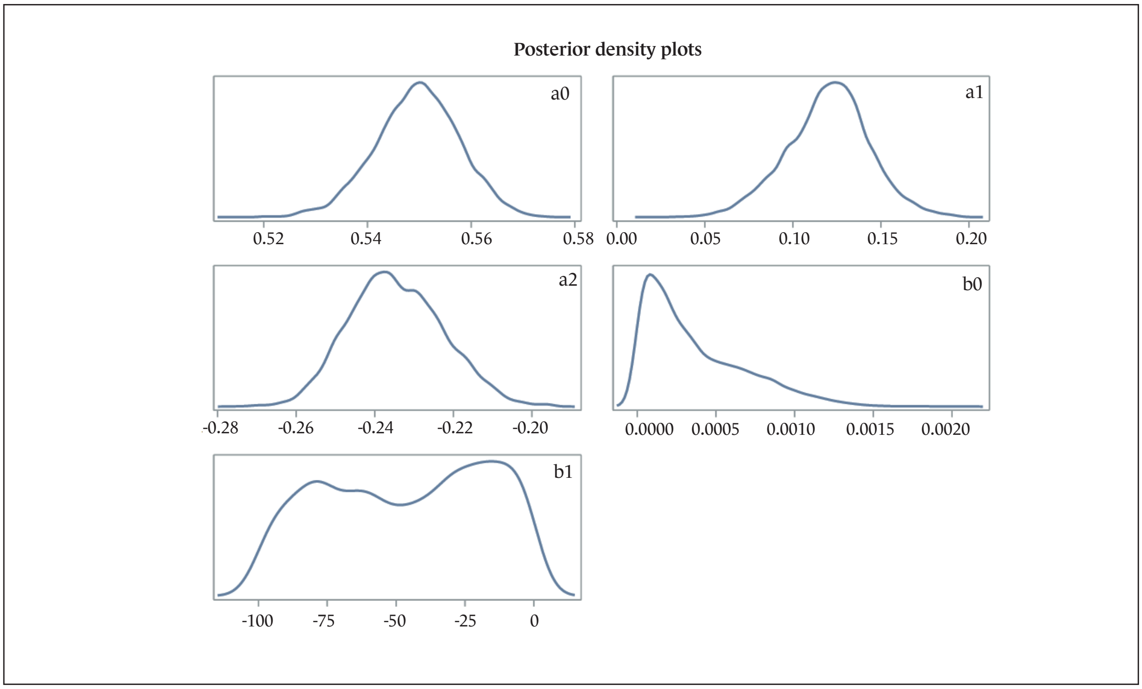
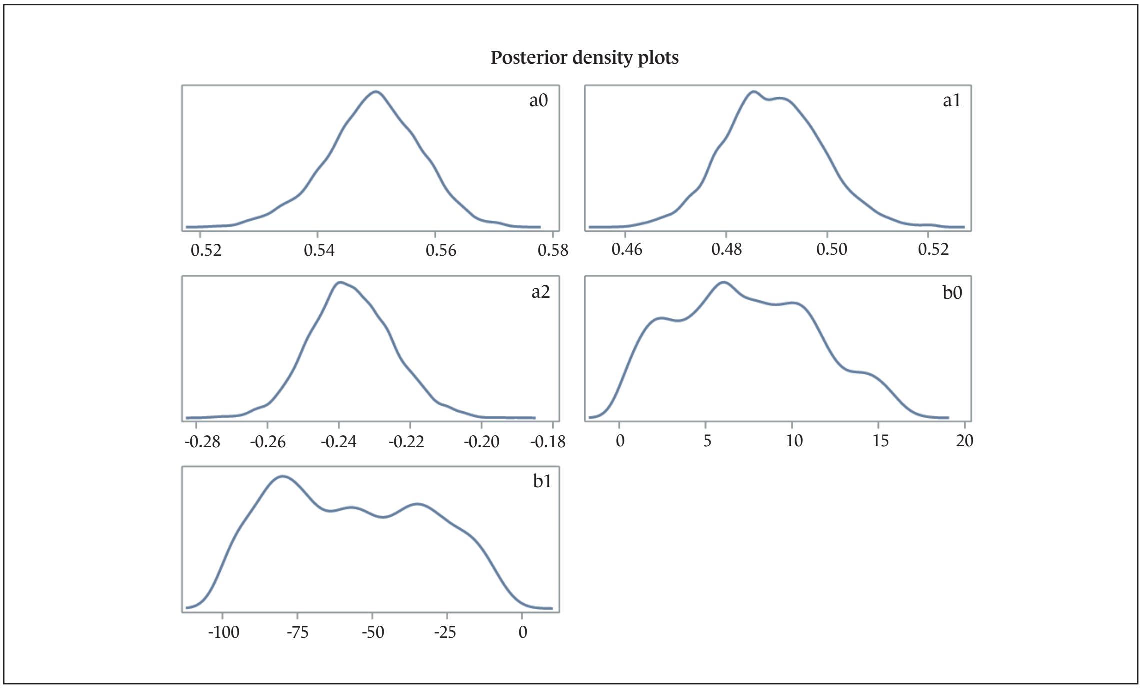


Figure 7
F_balance_ratio



The most meaningful for LGD estimation is current Loan to Value and it's dynamics. The proportion of granted loan and current loan (part already paid) compared to the value of collateral informs about the risk of recovering the remaining part. The significance of days and amount past due is also quite obvious. More days past due in the recent period increases the risk of credit loss. A higher balance ratio decreases potential credit loss.

The conclusions are given in the next section.

6. Summary and future research

The basic conclusion from the paper is that the combination of the classical approach with the Bayesian approach is possible and gives intuitive and promising results. In addition, it can be successfully applied to small samples on the example of LGD estimation and for unresolved cases. Using two binary models to predict two LGD modes close to 0 or 1 is a good idea to include more information to predict the LGD and differentiate the customer risk profile associated with these two groups (see also Ptak--Chmielewska, Kopciuszewski 2021). These two models were then established as input parameters to the Bayesian model. The biggest challenge and the most important point in the modelling procedure was the idea of using both resolved and unresolved cases in one LGD Bayesian model. This helped to avoid a model bias as can be seen in the classical approach, where unresolved cases are used for the later stage of the modelling step to adjust the final results. The approach of using unresolved cases in one model with resolved ones can be compared in its concept to using reject cases in the reject inference problem where the default status is unknow in the same way that future recoveries are unknown here. The final model quality was measured by R2. The discriminatory power of the model could only be checked for the resolved cases. The best quality is achieved for the model with f_balance_ratio as the explanatory variable. R2 is around 24% and proves the good model quality in comparison with the results in the LGD literature (Zhang, Thomas 2012; Matuszyk, Mues, Thomas 2010). In addition, looking at the posterior distribution of the model parameters, it seems that the f_balance_ratio is the best choice as well because the variance for parameters is not large and the distribution is quite smooth, symmetric and not concentrated on the border of the interval.

The promising results have also inspired us to deep dive into future research on unresolved cases and small sample LGD measurements with the application of more advanced methods.

References

Andritzky J. (2005), Default and recovery rates of sovereign bonds: a case study of the Argentine crisis, *Journal of Fixed Income*, 7, 97–107.

Anolli M., Beccalli E., Giordani T. (2013), *Retail Credit Risk Management*, Palgrave MacMillan, DOI: 10.1057/9781137006769.

Baesens B., Roesch D., Scheule H. (2016), *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*, John Wiley & Sons.

Bakshi G., Madan D., Zhang F. (2001), *Investing the sources of default risk: lessons from empirically evaluating credit risk models*, Working Paper, University of Maryland.

BCBS (2017), *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16)*, Basel Committee on Banking Supervision, https://eba.europa.eu/documents/10180/2033363/Guidelines+on+PD+and+LGD+estimation+%28EBA-GL-2017-16%29.pdf.

Belotti T., Crook J. (2007), Modelling and predicting loss given default for credit cards, *Quantitative Financial Risk Management Centre*, 28(1), 171–182.

Bijak K., Thomas L.C. (2015), Modelling LGD for unsecured retail loans using Bayesian methods, *Journal of the Operational Research Society*, 66(2), 342–352.

Brown I. (2012), *Basel II Compliant Credit Risk Modelling*, University of Southampton.

Brumma N., Urlichs K., Schmidt W.M. (2014), *Modeling downturn LGD in a Basel framework*, http://ssrn.com/abstract=2393351.

Chalupka R., Kopecsni J. (2009), Modelling bank loan LGD of corporate and SME segments: a case study, *Czech Journal of Economics and Finance*, 59(4), 360–382.

Christensen E., Henrik J. (2006), *Joint default and recovery risk estimation: an application to CDS data*, Working Paper, Copenhagen Business School.

Düllmann K., Trapp M. (2004), *Systematic risk in recovery rates – an empirical analysis of U.S. corporate credit exposures*, Working Paper, Deutsche Bundesbank.

Frye J. (2000), Depressing recoveries, *Risk*, 13/11, 108–111.

Frye J., Jacobs Jr. M. (2012), Credit loss and systematic loss given default, *The Journal of Credit Risk*, 8/1, 1–32.

Giese G. (2005), The impact of PD/LGD correlations on credit risk capital, *Risk*, April, 79–84.

Giese G. (2006), A saddle for complex credit portfolio models, *Risk*, 19/7, 84–89.

Hillebrand M. (2006), Modeling and estimating dependent loss given default, *Risk*, September, 120–125.

Huang X., Oosterlee C. (2011), Generalized beta regression models for random loss given default, *The Journal of Credit Risk*, 7(4), DOI: 10.21314/JCR.2011.150.

Hurt L., Felsovaly A. (1998), Measuring loss on Latin American defaulted bank loans, a 27-year study of 27 countries, *The Journal of Lending and Credit Risk Management*, 80, 41–46.

Izzi L., Oricchio G., Vitale L. (2012), *Basel III Credit Rating Systems*, Palgrave MacMillan, DOI: 10.1057/9780230361188.

Kosak M., Poljsak J. (2010), Loss given default determinants in a commercial bank lending: an emerging market case study, *Journal of Economics and Business*, 28(1), 61–88.

Loterman G., Brown I., Martens D., Mues C., Baesens B. (2012), Benchmarking Regression Algorithms for Loss Given Default Modelling, *International Journal of Forecasting*, 28(1), 161–170.

Luo X., Shevchenko P. (2013), Markov chain Monte Carlo estimation of default and recovery: dependent via the latent systematic factor, *Journal of Credit Risk*, 9(3), 41–76.

Matuszyk A., Mues Ch., Thomas L.C. (2010), Modelling LGD for unsecured personal loans: decision tree approach, *Journal of the Operational Research Society*, 61, 393–398.

Nielsen M., Roth S. (2017), *Basel IV: The Next Generation of Risk Weighted Assets*, John Wiley & Sons.

Papouskova M., Hajek P. (2019), Two-stage consumer credit risk modelling using heterogeneous ensemble learning, *Decision Support Systems*, 118, 33–45, DOI: 10.1016/j.dss.2019.01.002.

Ptak-Chmielewska A., Kopciuszewski P. (2021), Incorporating small-sample defaults history in loss given default models, *Journal of Credit Risk*, 17(4), 101–119, DOI: 10.21314/JCR.2021.009.

Ptak-Chmielewska A., Kopciuszewski P. (2023), Application of the Bayesian approach in loss given default modelling, *Bank i Kredyt*, 6.

Ptak-Chmielewska A., Kopciuszewski P., Matuszyk A. (2023), Application of the kNN-based method and survival approach in estimating loss given default for unresolved cases, *Risks*, 11, 42, DOI: 10.3390/risks11020042.

Pykhtin M. (2003), Unexpected recovery risk, *Risk*, 16, 74–78.

Tasche D. (2004), *The single risk factor approach to capital charges in case of correlated loss given default rates*, Quantitative Finance Papers, DOI: 10.48550/arXiv.cond-mat/0402390.

Van Berkel A., Siddiqi N. (2012), *Building loss given default scorecard using weight of evidence bins*, SAS Global Forum, https://support.sas.com/resources/papers/proceedings12/141-2012.pdf.

Yao X., Crook J., Andreeva G. (2017), Enhancing two-stage modelling methodology for loss given default with support vector machines, *European Journal of Operational Research*, 263(2), 679–689, DOI: 10.1016/j.ejor.2017.05.017.

Zieba P. (2017), Methods of extension of databases used to estimate LGD parameter, *Studia i Prace Kolegium Zarzadzania i Finansów*, 150, 31–55.

Zhang J., Thomas L.C. (2012), Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD, *International Journal of Forecasting*, 28, 204–15.

Appendix

The following logistic regression models were developed (see also Ptak-Chmielewska, Kopciuszewski 2023).

Table 1
The model predicting LGD = 0

Parameter	Estimate	p-value
Intercept	-2.9371	< 0.0001
The limit breach flag	-0.8732	0.0008
Change of residence flag	3.689E-7	0.0318
The number of employment months	-7.76E-7	0.0052
3 months savings	1.404E-6	0.0084
Ratio of 6 months savings to 3 month savings	-1.07E-6	0.0453
Refinancing flag	-2.6757	0.0003
Initial Loan to Value (LTV)	-2.8906	< 0.0001
LTV dynamics (3 months to 6 months)	0.9741	0.0021
Past due amount dynamics (3 months to 6 months)	1.299E-6	0.0030

Table 2
The description of variables for the model predicting LGD = 0

Variable	Description
The limit breach flag	Binary flag informing whether the customer has exceeded the limit
Change of residence flag	Binary flag informing whether the customer changed the residence
The number of employment months	The number of employment months
3 months savings	Savings from the last 3 months
Ratio of 6 months savings to 3 month savings	The ratio of the last 6 months savings to the last 3 month savings
Refinancing flag	Binary flag informing about the refinancing the exposure
Loan to Value (LTV)	LTV calculated in the application process
LTV dynamics (3 months to 6 months)	Ratio of LTV calculated 3 months ago and LTV calculated 6 months ago
Past due amount dynamics (3 months to 6 months)	Ratio of past due amount in the last 3 months and the last 6 months

Table 3

The quality measures for the model predicting LGD = 0

	Association of predicted probabilities and observed response		
Percent concordant	77.6	Somers' D	0.552
Percent discordant	22.4	Gamma	0.552
Percent tied	0.0	Tau-a	0.065
Pairs	204 750	c	0.776

Table 4

The model predicting LGD = 1

Parameter	Estimate	Wald Chi-Square	Pr > ChiSq
Intercept	-4.0969	159.5913	< 0.0001
Cover value after Household Prices Index (HPI)	-2.68E-6	16.2198	< 0.0001
Outstanding dynamics (last 6 months)	9.43E-7	17.8508	< 0.0001
Industry (Weight of Evidence (WoE) grouped)	0.1252	123.4599	< 0.0001
Life insurance flag	0.8718	16.5867	< 0.0001
LTV current	3.0650	102.2729	< 0.0001

Table 5

The description of variables for the model predicting LGD = 1

Variable	Description
Cover value after HPI	Collateral value index with HPI
Outstanding dynamics (last 6 months)	Dynamics of outstanding for the last 6 months
Industry (WoE grouped)	Customer industry transformed with WoE
Life insurance flag	Binary flag indicates whether the customer has life insurance
LTV current	Current LTV

Table 6

The quality measures for the model predicting LGD = 1

Association of predicted probabilities and observed responses			
Percent concordant	77.8	Somers' D	0.557
Percent discordant	22.2	Gamma	0.557
Percent tied	0.0	Tau-a	0.215
Pairs	672 060	c	0.77

Zastosowanie podejścia bayesowskiego z wykorzystaniem rozkładu beta do modelowania strat kredytowych

Streszczenie

Jednym z parametrów wyliczanych przez banki stosujące zaawansowane podejście w wewnętrznych ratingach (*advanced internal rating based* – AIRB) jest strata wynikająca z niewykonania zobowiązania (*loss given default* – LGD).

Podejścia do szacowania straty wynikającej z niewykonania zobowiązania mogą być różne. Podejście preferowane zarówno przez nadzorcę, jak i naukowców opiera się na wielkości odzysków zobowiązań kredytowych. W takim podejściu najbardziej problematyczne jest uzyskanie informacji o zdarzeniach niewykonania zobowiązań. W przypadku niektórych portfeli kredytowych (np. hipotecznych) liczba takich zdarzeń jest bardzo ograniczona. Wyniki uzyskane z zastosowania modeli statystycznych do kalkulacji parametru LGD będą obciążone małą liczebnością próby. Kolejnym wyzwaniem przy szacowaniu parametru LGD jest uwzględnienie przypadków niezakończonych jeszcze procesów odzysku.

Do estymacji LGD często stosuje się regresję i modele łączone, takie jak regresja liniowa i logistyczna. Zaproponowany w artykule model estymacji LGD opiera się na podziale odzysków na dwie klasy: wartości bliskie 0 lub wartości bliskie 1. Finalny model jest więc kombinacją dwóch modeli regresji binarnych.

Tradycyjne metody estymacji LGD nie uwzględniają jednak podejścia bayesowskiego, wykorzystującego informację *a priori*. Celem tego badania jest wykorzystanie podejścia bayesowskiego uwzględniającego założenia rozkładu beta dla niezakończonych procesów odzysku.

Proponowana metoda bierze pod uwagę specyfikę danych dla LGD zarówno w przypadku dwumodalnego rozkładu, jak i niepewności wynikającej z niezakończonych procesów odzysku. Prowadzi to do redukcji obciążeń modelu i bardziej precyzyjnych oszacowań. Połączenie podejścia klasycznego z metodologią bayesowską jest możliwe i prowadzi do wyników zgodnych z oczekiwaniami. Proponowane podejście może być stosowane w przypadku małych liczebnie prób danych. Największym wyzwaniem i jednocześnie najważniejszym punktem tej pracy było jednak wykorzystanie zarówno zakończonych, jak i niezakończonych procesów odzysku. W podejściu klasycznym przypadki niezakończonych procesów odzysku są wykorzystywane na dalszych etapach modelowania do skorygowania finalnych wyników estymacji. Niewątpliwą zaletą badania jest również prezentacja proponowanego podejścia do modelowania LGD na rzeczywistych danych o portfelu kredytów hipotecznych dla długiego okresu.

Słowa kluczowe: małe próby, LGD, podejście bayesowskie, rozkład beta, niezakończony proces odzysku