Bank i Kredyt 54(6), 2023, 625-650

BANK I KREDYT
BANK & CREDIT

Application of the Bayesian approach in loss given default modelling

Aneta Ptak-Chmielewska*, Paweł Kopciuszewski#

Submitted: 19 July 2023. Accepted: 27 September 2023.

Abstract

In some credit portfolios the number of observed defaults is always very limited. This is particularly evident in the Loss Given Default (LGD) estimation based on the new definition of default (the new definition of default was introduced in European banks in recent years) where only a small sample of empirical data is observed. The basic proposed LGD model is based on splitting recoveries into two classes of recoveries: value close to 0 or close to 1. This paper addresses also the problem with unresolved cases using the Bayesian approach, which assumes a distribution of further recoveries for unresolved cases. The Bayesian approach is considered with a combination of two binary models. The modelling approach for LGD is illustrated on real data for a long time period for mortgage loans. The proposed methodology takes into account the specificity of LGD data for both bimodal LGD distribution and uncertainty about unresolved cases, which lead to reduce a model bias.

Keywords: small samples, LGD, Bayesian approach, logistic and linear regression, unresolved cases

JEL: C1, C11

* Warsaw School of Economics; e-mail: aptak@sgh.waw.pl; ORCID: 0000-0002-9896-4240.
# Vistula University of Warsaw, ING Hubs Poland.

1. Introduction and background

Small samples in the LGD[1] estimation are always a challenge for researchers. In some portfolios like mortgages loan portfolios, the number of observed defaults is always very limited. In such a situation, all statistical models based on large samples will not work properly. The estimates will always be biased due to small sample properties. This is particularly evident in the LGD estimation based on the new definition of default (NDoD) where only a small sample of empirical data is observed. It is worth mentioning that the NDoD introduced the relative and the absolute thresholds for the purposes of days past due (DPD) counting. Other less important changes introduced by the NDoD are listed in the EBA GL (Guidelines on the application of the definition of default 2013).[2] The second approach in LGD modelling is to adequately account for the censored, unresolved exposures in the portfolio. Unresolved exposures are the exposures related to unfinished recovery processes at the moment of data generation. On the other hand, resolved cases are the exposures related to completed recovery processes at the moment of data generation for which all recoveries are known.

This paper proposes a modelling methodology to solve these two fundamental issues.

The auxiliary LGD model presented in this paper is based on splitting recoveries into two classes of recoveries: close to 0 or close to 1. This general observation of LGD distributions leads to the construction of an LGD model with the combination of two binary models. The second challenge when building the LGD model is to use unresolved cases in the estimation process. We address this problem using the Bayesian approach, which assumes a distribution of further recoveries for unresolved cases. The Bayesian approach is being considered with the LGD estimation for closed cases with a combination of two binary models.

The modelling approach for the LGD parameter is illustrated on real data for a long time period for mortgage loans. The proposed methodology takes into account the specificity of LGD data for both bimodal LGD distribution and uncertainty about unresolved cases to reduce a model bias.

The structure of the paper is the following. First, we provide a review of the existing literature both on the subject of LGD estimation, as well as literature discussing the Bayesian approach for LGD estimation. The second section discusses the justification of the choice of modelling approach, comparing the proposed approach with the classical one. The third section contains assumptions of the Bayesian approach for LGD modelling. The fourth section provides a data description and preliminary results of the two binary models. The final, fifth section, provides the main model estimation results. Finally, the conclusions of our results and some suggestions for further research are presented.

It is worth mentioning several paragraphs in the European Banking Authority (EBA) Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures[3] related to LGD modelling. Some of them are more important as they relate to unresolved cases, which is the focal point of the model covered by the proposed methodology.

[1] LGD is defined as a fraction of economic loss in the exposure of default (EAD).

[2] Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No. 75/2013, https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1597103/004d3356-a9dc-49d1-aab1-3591f4d42cbb/Final%20Report%20on%20Guidelines%20on%20default%20definition%20%28EBA-GL-2016-07%29.pdf?retry=1.

[3] 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.

- "However, to obtain a realistic value of long-run average LGD, the incomplete recovery processes should be included with future recoveries that are expected to be realized. The value of future recoveries is not an objective, observed measure but has to be estimated based on the recoveries factually observed on those cases that are already closed. As a result, the 'long-run average LGD' will also be a measure that is not fully objective, as it contains components that are estimated."
- "Institutions should obtain the long-run average LGD by adjusting the observed average LGD taking into account the information related to processes that were not closed ('incomplete recovery processes') and where the time from the moment of default until the moment of estimation is shorter than the maximum period of the recovery process specified for this type of exposures."
- "Institutions should analyse also other potential risk drivers that might become relevant after the date of default, including in particular the expected length of the recovery process and the status of the recovery process. Institutions should use the values of risk drivers as well as the values of collateral adequate to the reference dates specified in accordance with paragraphs 171 to 174."

The above paragraphs say that all the expected recoveries in the future should be taken into account when building the LGD model and that the proposed risk drivers should at least take into account:

– the expected length of the recovery process,

– the status of the recovery process.

The regulator also mentioned that the LGD adjustment should be done by including the information from incomplete processes. Our approach is in line with the requirement of taking into account the time in the default, which is correlated with the expected length of the recovery process but the adjustment of the LGD is changed to building only one model including unresolved cases at the same level in the modelling process as resolved cases.

There are many formulas and different approaches for LGD modelling in the literature. The best known were listed in Table 1. Many of them are based on Vasicek distribution or a similar approach.

Considering the classical approach for LGD modelling we have first of all averages computed from homogenous groups (Izzi, Oricchio, Vitale 2012), next go linear regression (Anolli, Beccalli, Giordani 2013; Loterman et al. 2012) or beta regression (Huang, Oosterlee 2011). This classical approach is of course preferred by authorities because they are clear and understandable for stakeholders. Recently more and more often new more advanced approaches including non-parametric are preferred by researchers as they are giving more precise and promising results. Unfortunately those "black--box" methods are not preferred by regulatory bodies. Quite innovative and interesting applications include decision trees (Belotti, Crook 2007), neural networks (Brown 2012) and Markov chains (Luo, Shevchenko 2013), as well as scoring-based methods (Van Berkel, Siddiqi 2012) and two-stage models (Brown 2012; Yao, Crook, Andreeva 2017; Papouskova, Hajek 2019).

The problem with unresolved cases has been discussed in the literature by Dermine and Neto de Carvalho (2006), Bastos (2010), and also Rapisarda and Echeverry (2013). Those papers present the application of both completed and unresolved cases to estimate recovery rate and exposure--weighted recovery rate curves. Additionally the exposure-weighted Kaplan-Meier estimator (Rapisarda, Echeverry 2013) has been used and modified to a default-weighted estimator. The most typical approach to include incomplete cases in modelling is treating those cases as complete (Baesens, Roesch, Scheule 2016). This approach may lead to significantly different estimates of real LGD values. Also the relationship between recovery rates and economic cycles has been explored by extrapolating recoveries

for the final LGD estimation for unresolved cases (Brumma, Urlichs, Schmidt 2014). For any method, under Basel IV, discussion of the assumptions regarding the treatment of unresolved cases should be also provided (Nielsen, Roth 2017). The need to recognize the impact of unresolved cases is clear from research that shows the potentially long time period before full recoveries are achieved. Some results based on different samples and portfolios shows that it may take up to four years from default to full recovery (Kosak, Poljsak 2010; Hurt, Felsovaly 1998).

Another approach for LGD estimation is a market-based LGD approach. For academics and researchers, the data are relatively accessible for different financial instruments. However, internal bank modelling are usually based on realized LGD, known as workout LGD, using historical recoveries and workout data. This is also the case of this paper. Implied LGD has been examined by some researchers using data from different countries and markets like BBB rated US corporate bonds (Bakshi, Madan, Zhang 2001) as well as for Argentinian government bonds (Andritzky 2005). Unfortunately results for market-derived LGD for corporate bonds characterize high estimation errors and low precision (Christensen, Henrik 2006; Pan, Singleton 2005).

Literature that studies the implications of small samples for loss (LGD) estimation is rather not very popular. Potential solutions proposed to overcome the small sample issue include the use of external databases, different time period criteria, or extrapolating future recoveries. One of the solutions was proposed by Chalupka and Kopecsni (2009), who used methods based on realized losses that can be used to extend small samples. They applied this solution for the estimation of LGD for the small and medium enterprise (SME) sector of the Czech Republic. The methods proposed include limiting the recovery time period or assuming a full recovery based on a proportion of the exposure. Zieba (2017) provides an overview of methods of increasing sample size. His work is based on using real LGD data, and this analysis supports the use of extrapolation of recovery rates as the most efficient way of increasing sample size for improving precision of LGD estimation.

Examples of Loss Given Default (LGD) estimations using a Bayesian approach are also very limited in the literature. One of the examples is LGD estimation for unsecured retail loans, because those estimations are often found difficult to model. The typical approach is two-step approach: there are two separate regression models that are estimated independently. However this approach can be problematic because it must be combined together to make the final predictions about LGD. In such a situation, LGD can be modelled using Bayesian methods (Bijak, Thomas 2015). The advantage in this situation is that only a single hierarchical model can be estimated instead of two separate models, making this a more appropriate approach. The authors of the mentioned paper used Bayesian methods, and the frequentist approach as a comparison, and applied them to the data on personal loans provided by one of the large UK banks. The posterior estimates of means of parameters that have been calculated by authors using the Bayesian approach appear to be very similar to the ones calculated in the frequentist approach. The main advantage of using the Bayesian model was an individual predictive distribution of LGD for each loan. Applications of such distributions also include so-called down-turn LGD calculations and the so-called stressed LGD calculations. Example of such application can be found in paper Jobst, Kellner and Rösch (2020). Authors developed and apply a Bayesian model for the loss rates given defaults (LGDs) of European sovereigns. Their approach comprises parameter risk and generates LGD forecasts under both regular and downturn conditions. With sovereign--specific rating information, they found that average LGD estimates vary between 0.46 and 0.64, while downturn estimates are between 0.50 and 0.86.

The biggest advantage of the approach proposed in this paper is the utilization of a Bayesian approach to avoid the need for aggregation of two different models results. We propose the theoretical and methodological elaboration and application on real banking data.

2. The choice of the modelling approach

There are two main ideas in the modelling proposed in this paper. The first one is to build two binary models predicting LGDs of 0 and 1, the second one is the incorporation of a Bayesian approach into the modelling process, especially for the unresolved cases. The rationale for these two choices is as follows:

- LGD distribution is bimodal with modes close to 0 and close to 1. Therefore it is more efficient to build estimation procedure on two binary models than on a continuous one. This approach was presented in a paper by Ptak-Chmielewska and Kopciuszewski (2021). Why two binary models but not just one that can be used to predict both states of LGD? Theoretically the score from the one binary model could be used to classify 0, 1 and other LGD values. These two binary models differ in their explanatory variables. The obligor risk profile associated with the prediction of LGD equal to 0 cannot be used to predict LGD of 1 and vice versa with good quality. This observation is also valid for the data contained in this paper. The conclusion is that the combination of the two binary models gives a more effective LGD forecast across the entire LGD range. In addition, modelling of continuous variables based on linear or nonlinear regressions is less efficient than modelling binary variables, but these methods cannot be compared directly.
- The LGD model has to be built taking into account all exposures, including unresolved cases, which is additionally emphasized by EBA regulations. Building the model in a more classic way where all resolved cases are the population for model building, but unresolved cases are used for the final adjustment of the model, can lead to a bias of the model. The most desired method is to include all cases in one model formula, keeping in mind that unresolved cases are censored cases in the model. Additionally, future recoveries as well as historically observed recoveries can be treated as random variables with a distribution. According to this approach, unknown parameters of this distribution can be estimated but also prior information can be incorporated into the structure of the model. The open questions are: Which distribution to choose for LGD for unresolved cases? How to combine a classic approach with the Bayesian methodology?

  Figure 1 gives a view of a possible combination of the classic and Bayesian approaches.

The comparison between the classic and currently proposed approaches as well as their pros and cons are summarized in Tables 2 and 3.

2.1. Assumptions for the Bayesian LGD estimation

The basic assumption for this LGD model 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 contained in the distribution parameters to the model. As total recoveries for resolved cases are known but partially are known for the unresolved ones, the final formula have to be different for them. This is a different approach compared to the classic one, where the resolved cases are used to

derive the model formula but the unresolved ones are used to adjust the model developed in the first modelling step. The more important the currently proposed approach is, the larger the unresolved population is observed. The current purpose in the paper is to present this new approach and test it with real data. It is also clear that the direct inclusion of the entire population in the modelling process makes the final model less biased. The bias of this model is related not to splitting the population into two different segments and treating the unresolved cases differently, but to the lack of total knowledge of recoveries for them. The formula presented below can be extended to more sophisticated models, which is allowed by Bayesian methodology. Currently the 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. In general, the structure of the model allows to take into account many other pieces information, such as external knowledge of the modelled variable, relations between coefficients, distributions for all parameters included in the model and the extension the model structure towards any needs. The uniqueness of the presented approach lies in applying semi-Bayesian methodology and taking into account both the resolved and unresolved cases in one model formula.

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.

$LGD_{obs,\ i}$ is the observed LGD up to the end of the recovery process for resolved cases and up to the censored time for unresolved cases, assigned to the *i*-th loan in the data.

$LGD_{total,\ i}$ is the overall LGD up to the end of the recovery process for both the resolved cases and unresolved cases, assigned to the *i*-th loan in the data.

$LGD_{obs,\ i}$ is assumed to be a random variable normally distributed, assigned to the *i*-th loan in the data.

$$LGD_{obs,\ i} \sim N\left(LGD_{mean,i}, LGD_{\sigma}\right) \tag{1}$$

where the mean and the standard deviation are equal to the LGD_{mean} and LGD_{σ} respectively.

From the Bayesian point of view the above formula for LGD with a normal distribution can be treated as likelihood, i.e. a conditional distribution of the LGD observed under the parameters used in the model. The prior distribution for all parameters introduced in this section, i.e. a_0, a_1, a_2, b is assumed to be non-informative.

As defined at the beginning, LGD is a fraction of economic loss in the EAD. Exceptionally, it may be lower than 0 and greater than 1 but the main range of the parameter is the [0, 1] interval. Nevertheless, the assumed normal distribution for LGD given a very small standard deviation covers the above interval with sufficiently high probability, which is acceptable from a practical point of view and sometimes used in Bayesian models. Formally, a normal density should be limited to a finite interval, but without loss of generality we can omit this assumption.

$LGD_{add,\ i}$ is the predicted additional part of the future loss (it is not the LGD estimator) for unresolved cases, assigned to the *i*-th loan in the data.

Suppose that the overall LGD can be predicted from the normal linear model based on the two logistic models developed in the previous section as follows:

$$LGD_{total,i} = a_0 + a_1 \cdot LGD_{0,\ i} + a_2 \cdot LGD_{1,\ i} + \varepsilon_i \tag{2}$$

and the expected mean $E\left(LGD_{total,i}\right)$ equals the above linear formula with three parameters.

Suppose that $LGD_{add,i}$ can be predicted from the normal linear model based on the basic information on the time (t) spent in the default state, which is recommended by regulations:

$$LGD_{add,i} = b \cdot t_i + \varepsilon_i' \tag{3}$$

where t_i is the time (in days) spent in default up till the censored time moment, assigned to the i-th loan in the data.

Then the above three random variables can be combined into one formula:

$$LGD_{obs,i} = LGD_{total,i} + LGD_{add,i} \tag{4}$$

where $LGD_{add,i}$ equals 0 for resolved cases.

Next, the mean and the standard deviation for the observed LGD of the i-th facility are defined separately for the resolved and unresolved cases as follows:

- Resolved cases

$$LGD_{mean,i} = E\left(LGD_{total,i}\right) = a_0 + a_1 \cdot LGD_{0,i} + a_2 \cdot LGD_{1,i} \tag{5}$$

LGD_{σ} is a single parameter to estimate not involved in any formula dependent on other variables.

- Unresolved cases

$$LGD_{mean,i} = E\left(LGD_{total,i}\right) + E\left(LGD_{add,i}\right) = a_0 + a_1 \cdot LGD_{0,i} + a_2 \cdot LGD_{1,i} + b \cdot t_i \tag{6}$$

On the other hand, from the above formulas, the expected value of total LGD can be calculated conditionally given the observed part of LGD as follows:

$$E\left(LGD_{total,i}\right) = E\left(LGD_{obs,i}\right) - E\left(LGD_{add,i}\right) = LGD_{mean,i} - E\left(LGD_{add,i}\right) \tag{7}$$

Therefore, this formula can be also applied to predict the total LGD for unresolved cases given the observed part of the LGD up to time t as follows:

$$E\left(LGD_{total,i}\right) = LGD_{mean,i} - b \cdot t_i \tag{8}$$

It is the expected value of the observed LGD equal to the prediction of the total LGD adjusted for the expected prediction for the future recoveries.

Ultimately, the use of the above formulas based on a linear relationship and normal model requires the LGD winsorization. In general LGD can be lower than 0 in very exceptional cases and strongly justified by business rules. Values greater than 1 are related to the collection process with no recoveries

and increased costs. There were no LGD outliers in the development data, but this winsorization step should be kept for all other data for which the LGD formula is used. Hence the final estimator for the total LGD for performing cases is:

$$LGD_{mean,i} = \max\left(0, \min(1,\ a_0 + a_1 \cdot LGD_{0,i} + a_2 \cdot LGD_{1,i})\right) \tag{9}$$

and the final estimator for the total LGD for 10 unresolved cases is:

$$LGD_{mean,i} = \max\left(0, \min(1,\ a_0 + a_1 \cdot LGD_{0,i} + a_2 \cdot LGD_{1,i} + b \cdot t_i)\right) \tag{10}$$

2.2. The data and two binary models for LGD prediction

The data used for the modelling process comes from the years 2008 to 2018. It includes 1867 observations,[4] of which 314 are unresolved, and 400 explanatory variables. All unique loans for the customer are assigned to different reporting dates from the abovementioned period of time. The main assumption for the construction of the sample is that the reporting date is randomly selected for the loan from the 1st to the 12th month before entering to the default state. All explanatory variables are current as of the selected reporting date. The target variables of the two binary models are based on the realized economic loss of the loan.

It should be also added that the data was cleaned before being used for the modelling purpose. The trends over time of the target mean variables[5] $LGD_i = 0$ and $LGD_i = 1$ are shown in Table 4 and Figures 2 and 3.

The changing trend of the share of LGD equal to zero or one should not be interpreted as a change in the population pattern, because it is biased with the shorter observation window for recent years. On the other hand, it can be seen that the share of $LGD_i = 1$ in the population is stable over time (excluding recent years with a small sample). The variables admitted for the modelling process were pre-selected initially by business experts and filtered to those with the best properties according to the stability Population Stability Index (PSI) and Gini measures. While building logistic regression, the stepwise method was used for the later selection and applied along with the modeler evaluation. This additional assessment was focused mainly on checking the good properties of the model such as the same sign at the variable as in the univariate analysis, not so strong correlation between variables and the significance tests for model parameters.

The general formula for the both models is:

$$P^l\left(LGD_i = l\right) = \frac{1}{1 + \exp[-(\beta_0^l + \sum_{j=1}^{k_l} \beta_j^l \cdot x_{i,j}^l)]},\ l \in \{0,1\} \tag{11}$$

$$P^l\left(LGD_i \neq l\right) = 1 - P^l\left(LGD_i = l\right)$$

[4] Observation is the unique loan of the customer, in the case of several products per customer, a random one is selected to the sample.

[5] LGD is a continuous variable from interval [0–1] with bimodal distribution. Two values, zero and one, are its modes with positive probability.

where:

β_i^l – the l-th model parameters, $j = 0,\ldots, k_l$,

$x_{i,j}^l$ – the l-th model variables, $j = 1,\ldots, k_l$ for the i-th loan.

Both the models predicting $LGD_i = 0$ and $LGD_i = 1$ separately were built on separate populations and probabilities of the opposite events for these two models are defined above. The main reason for building separate models was to observe different risk profiles determined by different risk drivers used to predict the event of $LGD_i = 0$ and the event of $LGD_i = 1$.

Tables 5–10 show both the structure and quality of the developed models.

Summarizing the results presented in above mentioned tables:

- The quality of these two models is almost the same.
- Risk profile of the exposures that are predicted to be 0 is based mainly on behavioural variables:
  – initial LTV,
  – refinancing flag,
  – limit breach flag,
  – number of employment months,
  – LTV dynamics,
  – 3 months savings,
  – past due amount dynamics (3 months to 6 months).
- The most significant variables in the model LGD = 1 are:
  – industry,
  – LTV current.
- The common part of the models is LTV, but in the first model initial LTV is more predictive and in the second one, LTV current.

3. The main model results

Both binary LGD models and the final Bayesian model were built using SAS. The nlmixed procedure was used to build the final Bayesian LGD model. All dependencies and formulas described in the previous section were incorporated in this SAS procedure. It is worth mentioning that the Monte Carlo calculations performed in Bayesian modelling depend on the initial values of the algorithm, but changing them gave the same results, hence the conclusion is that the algorithm is stable on the data we used in Bayesian modelling. In Bayesian modelling final results are posterior estimators. Table 11 presents all of them. All parameters are statistically significant and the signs of the coefficients are intuitive.

An alternative formula for the LGD estimation was tested using an additional variable such as the amount of recoveries collected before the end of the observed process (up till time t). This later variable turned out to be insignificant, therefore the additional part of loss was defined only on the basis of the variable t.

The final LGD estimator predicting the total LGD_i for the i-th loan can be defined as follows:

$$LGD_{mean,i} = \max\left(0, \min\left(1, 0.2751 - 0.5594 \cdot LGD_{0,i} + 0.5980 \cdot LGD_{1,i}\right)\right) \tag{12}$$

and it can be used for performing cases.

Analysing the LGD prediction on the data used for modelling, the maximum and the minimum values of LGD are as follows: 0.5980 and -0.2843. Therefore winsorizing on the development data is needed only to the lower LGD value, but on the other hand it can indicate the symptoms of rising costs with no recoveries. Therefore the decision of the final winsorization should be made on the basis of the knowledge of the recovery process and confirmation that LGD greater than 1 are possible.

The additional part of loss for unresolved cases is defined as follows:

$$LGD_{add,i} = 0.00031 \cdot t_i \tag{13}$$

From the formula it results that the total LGD for in-default unresolved cases can be calculated as:

$$E\left(LGD_{total,i}\right) = \max\left(0, \min(1,\ LGD_{mean,i} - 0.00031 \cdot t_i\right)\right) \tag{14}$$

The two above formulas for LGD can be applied to predict LGD both for resolved and unresolved cases. In other words, they can be used for performing cases and in-default cases with known the time spent in default status until the LGD prediction is computed.

Tables 12 and 13 show the numeric results obtained from the SAS system based on the predicted LGD for closed cases.

The first analysis is the correlation between the following predicted LGDs and LGD realized (see Tables 12 and 13): LGD realized, LGD_{total}, LGD_{add}, $LGD_{add\_ratio} = \frac{LGD_{add}}{LGD_{total}}$ and two predicted LGD values from the binary models, i.e. LGD_i = 0 and LGD_i = 1. The analysis is divided into the resolved and unresolved cases.

The above results give some interesting information:

– the total LGD prediction is mainly correlated with the prediction that LGD_i = 1;

– both binary models predicting LGD_i = 0 and LGD_i = 1 are correlated at a lower level;

– the total LGD prediction is quite strongly correlated with the additional part of the loss for the unresolved cases; the lower the additional part of the loss, the higher should be the total predicted LGD.

The second analysis is to compare the basic statistic between the resolved and unresolved cases (see Tables 14 and 15).

The summary of the above results is intuitive and can be detailed below:

- LGD realized is much higher for the unresolved than for resolved cases, which is natural as recoveries are observed partially for the unresolved cases.
- The average of LGD_i = 0 prediction is almost the same for the resolved and unresolved cases.
- The average of LGD_i = 1 prediction is twice as high for the unresolved than for the resolved cases and the total LGD is also much higher for the unresolved cases. It indicates the greater losses for unresolved cases, which is also intuitive.
- The additional future share of losses is around 80% of the observed partial LGD based on recoveries up to the censored time *t*.
- Minimum and maximum values of LGDs without winsorization are within the range [0–1].

The last analysis is based on distributions for LGD realized, LGD_{total}, LGD_i = 0 estimator, LGD_i = 1 estimator and their comparison. It can be seen that *t* distributions for the LGD_0 have

a similar shape, but the distribution of the LGD_1 for unresolved cases appears to be univariate but for resolved cases it is unimodal (see Figures 4 and 5).

The conclusion from Figures 6 and 7 is that the variance of the distribution for the total predicted LGD is much higher for unresolved cases, but the observed LGD is concentrated close to 1 for unresolved cases and closer to 0 for resolved ones.

4. Conclusions and future research

The most important conclusion from the paper is that the combination of the classic approach with the Bayesian methodology is possible and gives intuitive results. In addition, it can be successfully applied to small samples on the example of LGD estimation. Using two binary models to predict two LGD modes close to 0 or 1, it is a good idea to include more information to predict the LGD and differentiate the customer risk profile associated with these two groups. 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 unknown as well as future recoveries are unknown here. Following this approach we can continue by focusing on:

– taking into account the other LGD distributions such as Beta distribution,

– extending the Bayesian methodology to informative priors, that can be discovered from data or previous experiences,

– inclusion of the more explanatory information in the part of the model related to unresolved cases,

– reducing the dimensionality of the Bayesian approach starting from the business assumptions and model properties required from the business point of view.

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, 15, University of Maryland.

Bastos J. (2010), Forecasting bank loans loss-given-default, *Journal of Banking and Finance*, 34(10), 2510–2517, DOI: 10.1016/j.jbankfin.2010.04.011.

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.

Dermine J., Neto de Carvahlo C. (2006), Bank loan losses-given-default: a case study, *Journal of Banking and Finance*, 30(4), 1219–1243.

Düllmann K., Trapp M. (2004), *Systematic risk in recovery rates – an empirical analysis of U.S. corporate credit exposures*, Discussion Paper Series 2, 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.

Jobst R., Kellner R., Rösch D. (2020), Bayesian loss given default estimation for European sovereign bonds, *International Journal of Forecasting*, 36(3), DOI: 10.1016/j.ijforecast.2019.11.004.

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.

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

Pan J., Singleton K. (2005), *Default and recovery implicit in the term structure of sovereign CDS spreads*, Working Paper, Stanford University.

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.

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

Rapisarda G., Echeverry D. (2013), A non-parametric approach to incorporating incomplete workouts into loss given default estimates, *Journal of Credit Risk*, 9(2), DOI: 10.21314/JCR.2013.159.

Tasche D. (2004), *The single risk factor approach to capital charges in case of correlated loss given default rates*, arXiv.org, DOI: 10.2139/ssrn.510982.

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.

Appendix

Table 1
The most frequently used formulas for LGD estimation

Author	Description	Formula
Frye-Jacobs (2012)	The LGD function connects the conditionally expected LGD rate (cLGD) to the conditionally expected default rate (cDR)	$\Phi\left[\Phi^{-1}\left[cDR\right]-k\right]/cDR$ $k = LGDrisk\,index = \left(\Phi^{-1}\left[PD\right]-\Phi^{-1}\left[EL\right]\right)/\sqrt{1-\rho}$
Frye (2000)	Recovery is a linear function of the normal risk factor associated to the Vasicek distribution	$1-\left(\mu+\sigma q\left(\sqrt{1-\rho}\,\Phi^{-1}\left[cDR\right]-\Phi^{-1}\left[PD\right]\right)/\sqrt{\rho}\right)$ μ – recovery mean, σ – recovery SD, q – recovery sensitivity
Pykhtin (2003)	Proposes parameterization of the amount, volatility, and systematic risk of a loan's collateral and infers the loan's LGD	$\Phi\left[\frac{-\frac{\mu}{\sigma}-\beta Y}{\sqrt{1-\beta^2}}\right]-\exp\left[\mu+\sigma\beta Y+\frac{\sigma^2}{2}\left(1-\beta^2\right)\right]\Phi\left[\frac{-\frac{\mu}{\sigma}-\beta Y}{\sqrt{1-\beta^2}}-\sigma\sqrt{1-\beta^2}\right]$ $Y=\left(\Phi^{-1}\left[PD\right]-\sqrt{1-\rho}\Phi^{-1}\left[cDR\right]\right)/\sqrt{\rho}$ μ – log recovery mean, σ – log recovery SD, β – recovery correlation
Tasche (2004)	Assumes a connection between LGD and the systematic risk factor at the loan level; the idiosyncratic influence is integrated	$\int_{-\Phi^{-1}[cDR]}^{\infty}\phi[z]\,BetaCDF^{-1}\left[\frac{\Phi\left[\sqrt{1-\rho}\,\Phi^{-1}[cDR]-\Phi^{-1}[PD]+\sqrt{1-\rho}z\right]-1+PD}{PD},\right.$ $\left.a=\frac{ELGD\,(1-v)}{v},\ b=\frac{(1-ELGD)(1-v)}{v}\right]dz/cDR$ *ELGD* – expected LGD, v = fraction of maximum variance of Beta distribution
Giese (2005)	Makes a direct specification of the functional form linking cLGD to cDR	$1-a_0\left(1-PD^{a_1}\right)^{a_2}$ a_1, a_2, a_3 – values to be determined
Hillebrand (2006)	Introduces a second systematic factor that is integrated out to produce cLGD given cDR	$\int_{-\infty}^{\infty}\Phi\left[a-\frac{bdc}{e}+\frac{bd}{e}\Phi^{-1}\left[cDR\right]-b\sqrt{1-d^2x}\right]\phi[x]dx$ a, b – parameters of *cLGD* in second factor, d – correlation of latent factors $c=\frac{\Phi^{-1}\left[PD\right]}{\sqrt{1-\rho}};\ e=\frac{\sqrt{\rho}}{\sqrt{1-\rho}}$
Giese (2006)	Uses the beta distribution on systematic factor *Y*	$LGD \sim Beta(\alpha(Y), \beta(Y))$
Düllmann, Trapp (2004)	The recovery rate is modelled as a logit transformation of a normally distributed random variable *Y*	$Y_j=\mu+\sigma\sqrt{\omega}X+\sigma\sqrt{1-\omega}W_j,\ X\sim N(0,1),\ W_j\sim N(0,1)$ $R\left(Y_j\right)=\frac{\exp\left(Y_j\right)}{1+\exp\left(Y_j\right)}$

Table 2
Comparison between the classic and currently proposed approaches

	Classic approach	Proposed approach
Population for model building	The population to build the main LGD model refers only to resolved cases. Unresolved ones are used at a later stage	The population includes both resolved and unresolved cases with recoveries observed up till the moment of data generation, but with additional recoveries being estimated for unresolved cases
Estimation procedure	The main model is built on the resolved cases and adjusted on the entire population with both resolved and unresolved. Mostly unresolved are used only at the adjustment stage after applying the dragging estimation algorithm for them	The first two binary regression estimate LGD = 0 and LGD = 1 events are built. Two estimators are used as input variables in the Bayesian model. Only one model is built on the entire population but a separate prediction is estimated for unresolved cases

Table 3
Pros and cons for the classic and currently proposed approach

Approach	Pros	Cons
Classic	Well-known, accepted procedures, simple results More real examples to benchmark	The main model built on the population is biased with excluding unresolved cases No possibility of including expert knowledge or random distribution of parameters Additional assumptions for regression models
Proposed	Population of the main model includes all cases (resolved and unresolved) Possibility of including external expert knowledge A wide range of distributions for parameters No restrictions on the complexity of the model	More complicated procedure, subjective assumptions on LGD distribution More complex calculations for the posterior parameters of a multivariate model or hierarchical model

Table 4
LGD = 0 and LGD = 1 shares in the population of both all and the resolved cases from 2008 to 2018

Year of default	The entire population			Resolved cases only		
	Avg ($LGD_i = 0$)	Avg ($LGD_i = 1$)	Population size	Avg ($LGD_i = 0$)	Avg ($LGD_i = 1$)	Population size
2008	0.055276	0.120603	199	0.055276	0.120603	199
2009	0.073171	0.209756	205	0.073171	0.209756	205
2010	0.048544	0.203883	103	0.048544	0.203883	103
2011	0.025000	0.166667	120	0.025641	0.162393	117
2012	0.059701	0.159204	201	0.063158	0.147368	190
2013	0.034335	0.171674	233	0.037037	0.162037	216
2014	0.064655	0.159483	232	0.072816	0.140777	206
2015	0.088542	0.182292	192	0.100592	0.142012	169
2016	0.085106	0.489362	141	0.151899	0.189873	79
2017	0.092857	0.614286	140	0.260000	0.000000	50
2018	0.059406	0.792079	101	0.315789	0.000000	19

Table 5
Model predicting $LGD_i = 0$

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 month savings	1.404E-6	0.0084
Ratio of 6 month 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 6
Description of variables for the model predicting $LGD_i = 0$

Variable	Description
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
Number of employment months	The number of employment months
3 months savings	Savings from the last 3 months
Ratio of 6 month savings to 3 month savings	The ratio of the last 6 month savings to the last 3 month savings
Refinancing flag	Binary flag informing about the refinancing the exposure
Initial 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 7
Quality measures for the model predicting $LGD_i = 1$

Association of predicted probabilities and observed response			
P concordant	77.6	Sommers' 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 8
Model predicting $LGD_i = 1$

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 9
Description of variables for the model predicting $LGD_i = 1$

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 10
Quality measures for the model predicting $LGD_i = 1$

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	672060	c	0.77

Table 11
Structure and statistics of the Bayesian model

Parameter	Estimate	Standard error	DF	T value	p-value	95% confidence limits	
a_0	0.2751	0.01761	1867	15.62	< 0.0001	0.2405	0.3096
a_1	-0.5594	0.13770	1867	-4.06	< 0.0001	-0.8295	-0.2894
a_2	0.5980	0.04388	1867	13.63	< 0.0001	0.5119	0.6841
sigma	0.1264	0.00414	1867	30.55	< 0.0001	0.1183	0.1346
b	0.00031	0.00002	1867	17.25	< 0.0001	0.0003	0.0003

Table 12
Correlation of the predicted LGDs and realized LGD for resolved cases

	Pearson correlation coefficients			
	LGD realized	$LGD_i = 0$	$LGD_i = 1$	LGD_{total}
LGD_{obs}	1.00000	-0.20301	0.21901	0.25277
$LGD_i = 0$	-0.20301	1.00000	-0.28360	-0.55562
$LGD_i = 1$	0.21901	-0.28360	1.00000	0.95488
LGD_{total}	0.25277	-0.55562	0.95488	1.00000

Table 13
Correlation of the predicted LGDs and realized LGD for unresolved cases

	Pearson correlation coefficients					
	LGD realized	$LGD_i = 0$	$LGD_i = 1$	LGD_{total}	LGD_{add}	LGD_{add_ratio}
LGD_{obs}	1.00000	-0.12036	0.46943	0.46485	-0.50372	-0.63900
$LGD_i = 0$	-0.12036	1.00000	-0.20827	-0.39593	-0.13966	0.13517
$LGD_i = 1$	0.46943	-0.20827	1.00000	0.98061	-0.44688	-0.68201
LGD_{total}	0.46485	-0.39593	0.98061	1.00000	-0.39158	-0.66741
LGD_{add}	-0.50372	-0.13966	-0.44688	-0.39158	1.00000	0.88434
LGD_{add_ratio}	-0.63900	0.13517	-0.68201	-0.66741	0.88434	1.00000

Table 14
Basic statistics for the predicted LGDs and realized LGD for resolved cases

Variable	No.	Mean	Std Dev	Minimum	Maximum
LGD_{obs}	1553	0.35315	0.38357	0.00036	1.85892
$LGD_i = 0$	1553	0.06197	0.06231	0.00034	0.48609
$LGD_i = 1$	1553	0.22457	0.16318	0.00116	0.95183
LGD_{total}	1553	0.37473	0.11254	0.01379	0.84318

Table 15
Basic statistics for the predicted LGDs and realized LGD for unresolved cases

Variable	No.	Mean	Std Dev	Minimum	Maximum
LGD_{obs}	314	0.93789	0.11749	0.15706	1.17600
$LGD_i = 0$	314	0.06612	0.05701	0.00018	0.28415
$LGD_i = 1$	314	0.44026	0.24987	0.02943	0.96378
LGD_{total}	314	0.50139	0.15915	0.16294	0.83770
LGD_{add}	314	0.33117	0.19006	0.11346	0.88722
LGD_{add_ratio}	314	0.80113	0.66583	0.15230	3.93828

Figure 1
Diagram of LGD model building

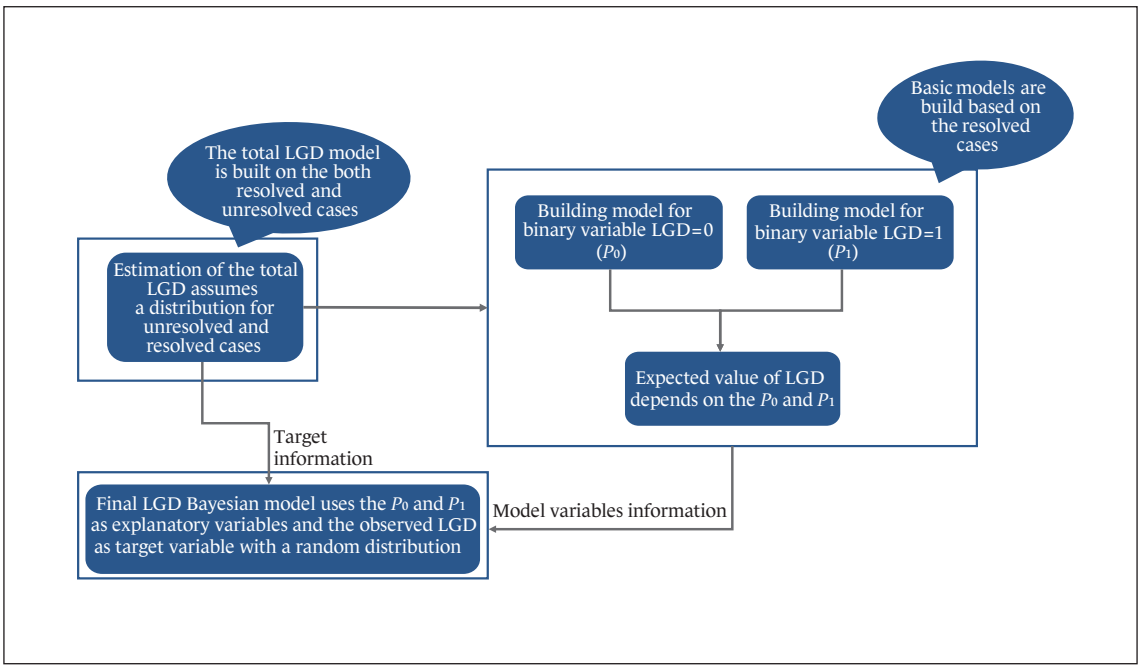


Figure 2
LGD = 0 and LGD = 1 shares in the population of the resolved cases

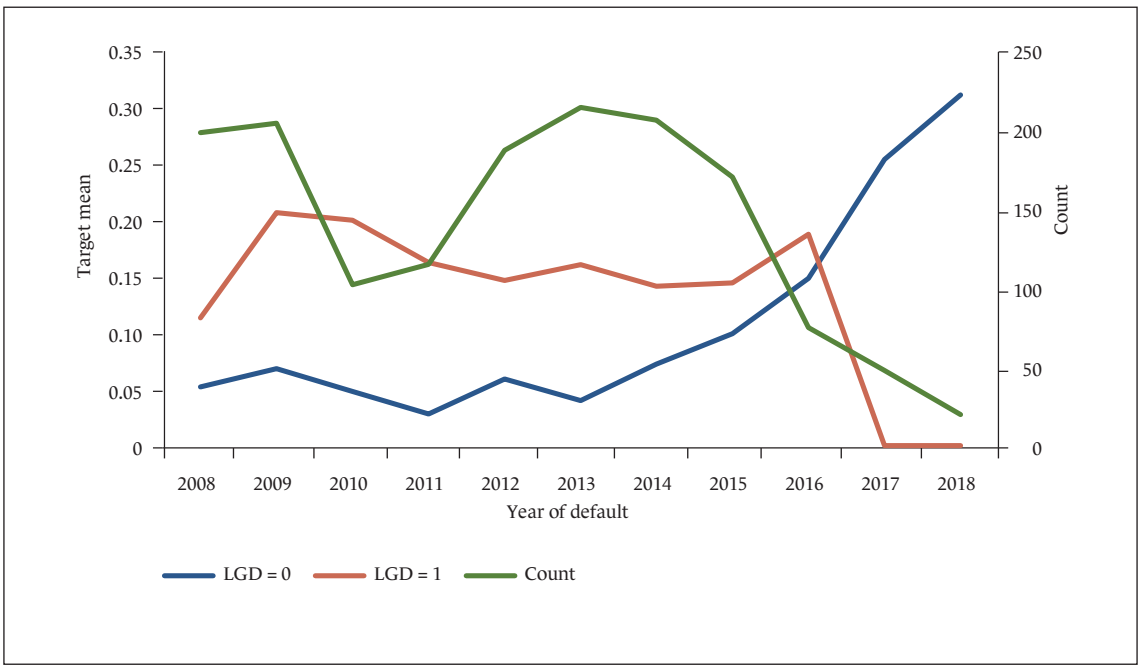


Figure 3
LGD = 0 and LGD = 1 shares in the population of all cases

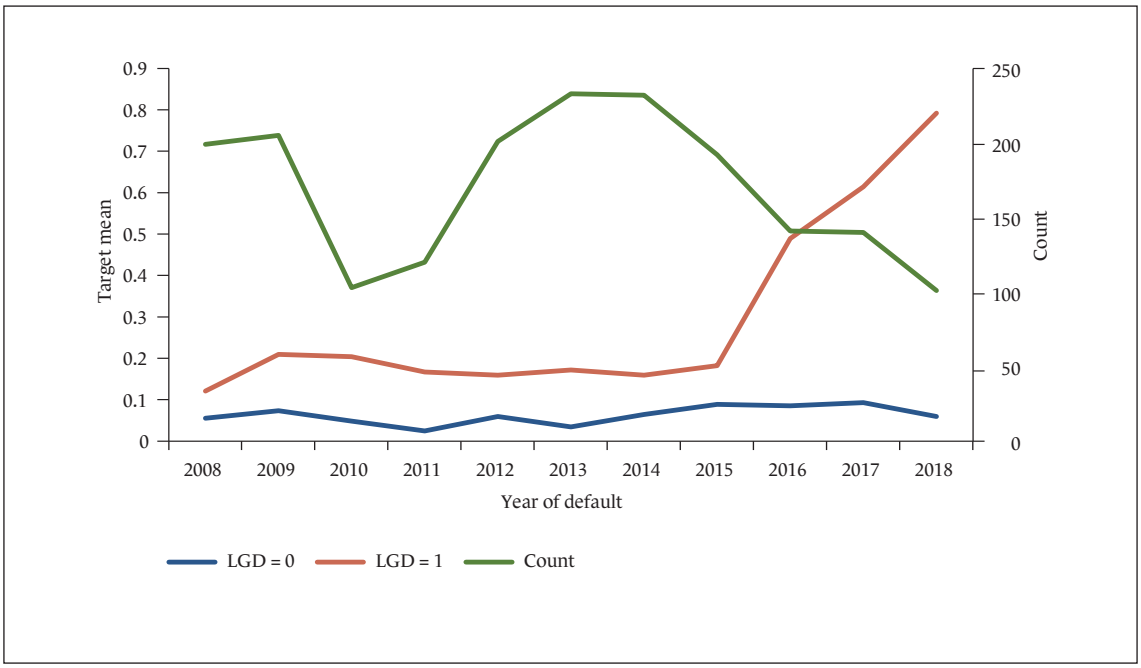


Figure 4
Distribution of $LGD_i = 0$ estimator divided into the resolved and unresolved cases

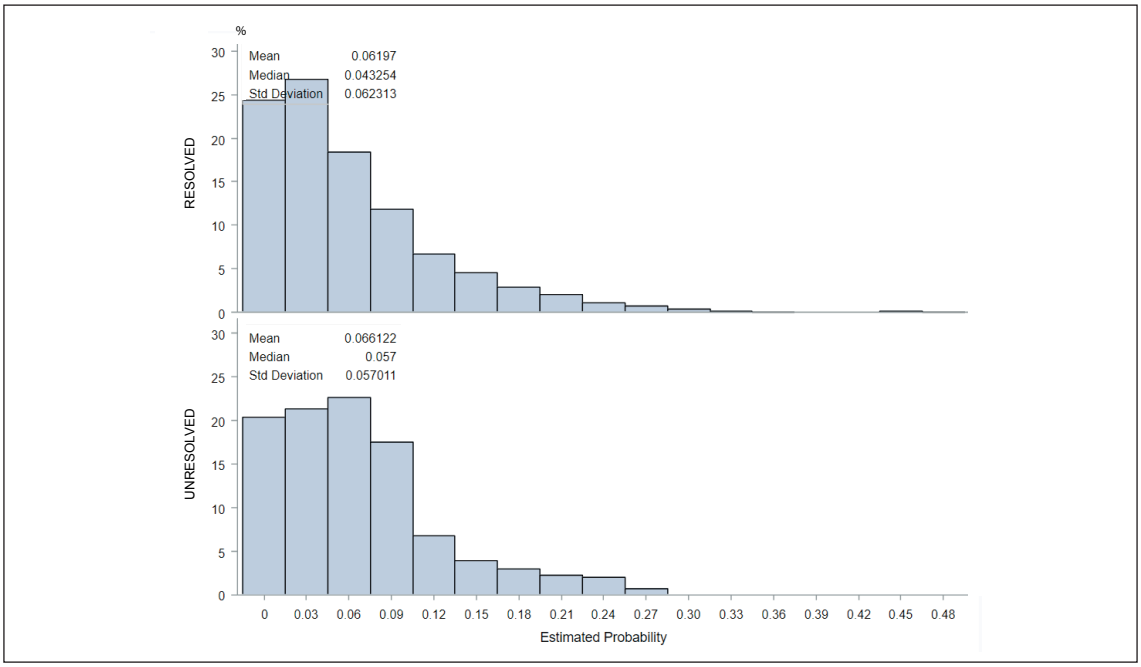


Figure 5
Distribution of $LGD_i = 1$ estimator divided into the resolved and unresolved cases

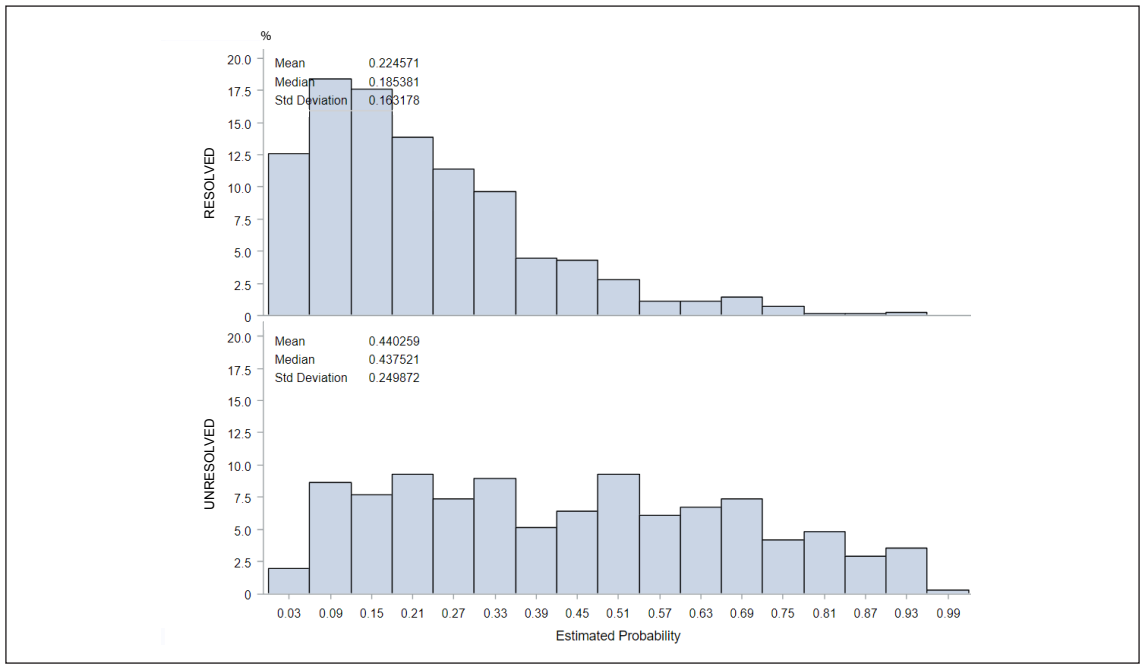


Figure 6
Distribution of the total predicted LGD divided into the resolved and unresolved cases

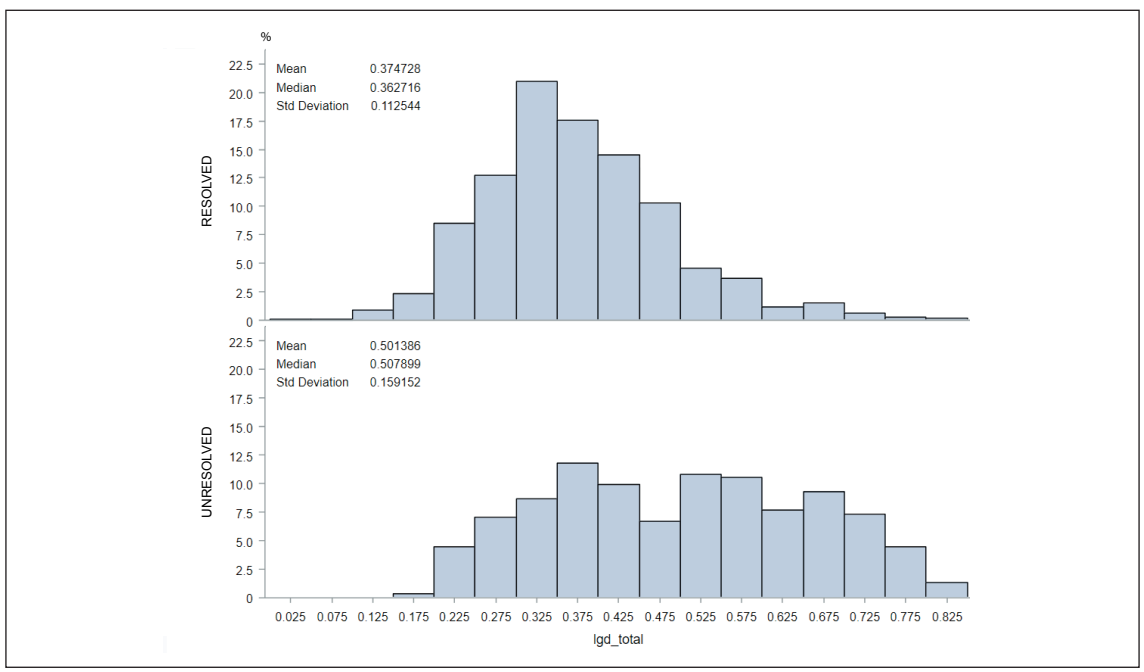
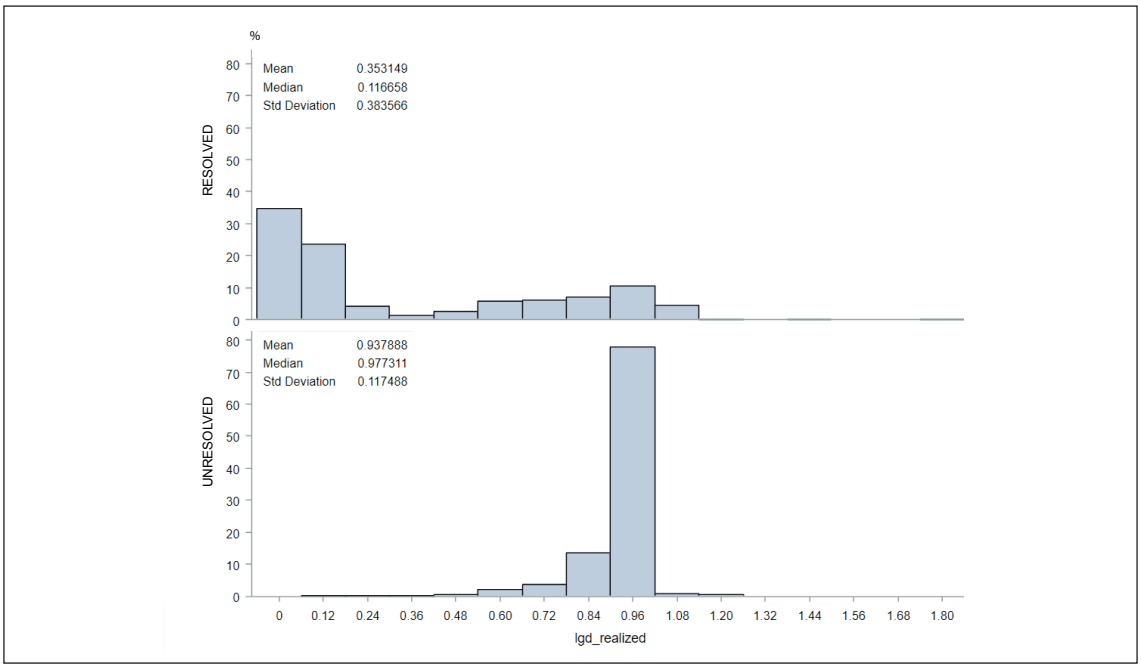


Figure 7
Distribution of the realized LGD divided into the resolved and unresolved cases



Zastosowanie podejścia bayesowskiego do modelowania strat wynikających z niewykonania zobowiązania

Streszczenie

Zgodnie z zaawansowanym podejściem w wewnętrznych ratingach (*advanced internal rating based* – AIRB) banki mogą samodzielnie wyliczać parametry ryzyka na podstawie własnych baz danych. Parametrami tymi są: prawdopodobieństwo niewykonania zobowiązania (*probability of default* – PD), ekspozycja w momencie niewykonania zobowiązania (*exposure at default* – EAD) oraz strata wynikająca z niewykonania zobowiązania (*loss given default* – LGD). Podejście do szacowania straty LGD, które preferują zarówno nadzorcy, jak i naukowcy, opiera się na wielkości odzysków zobowiązań kredytowych. W przypadku niektórych portfeli kredytowych, np. kredytów zabezpieczonych hipotecznie, liczba obserwowanych zdarzeń niewykonania zobowiązania jest jednak zawsze ograniczona. W takiej sytuacji modele statystyczne nie będą działały poprawnie. Oszacowania zawsze będą obciążone ze względu na małą liczebnie próbę danych.

Jest to szczególnie widoczne w przypadku szacowania strat kredytowych zgodnie z nową definicją niewykonania zobowiązania, gdy dostępne są tylko małe próby danych empirycznych. Podstawowy model estymacji LGD opiera się na podziale odzysków na dwie klasy: wartości bliskie 0 lub wartości bliskie 1. Prowadzi to do skonstruowania modelu będącego kombinacją dwóch modeli binarnych. Kolejnym wyzwaniem w procesie estymacji LGD jest uwzględnienie przypadków niezakończonych jeszcze procesów odzysku.

Tradycyjne metody estymacji LGD nie uwzględniają podejścia bayesowskiego. Zazwyczaj popularnym podejściem jest regresja i modele łączone, takie jak regresja liniowa i logistyczna. Celem tego badania jest wykorzystanie podejścia bayesowskiego uwzględniającego założenia o niezakończonych procesach odzysku. Podejście bayesowskie jest uwzględnione do estymacji LGD dla zakończonych przypadków odzysku w połączeniu z dwoma modelami binarnymi. Zaletą badania jest również to, że prezentuje proponowane podejście do modelowania LGD na rzeczywistych danych o portfelu kredytowym dla długiego okresu.

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.

Najważniejszym wnioskiem wynikającym z badań jest to, że połączenie podejścia klasycznego z metodologią bayesowską jest możliwe i prowadzi do wyników zgodnych z intuicją. Dodatkowo takie podejście może być stosowane w przypadku małych liczebnie prób danych. Zastosowanie dwóch modeli binarnych pozwala na wykorzystanie informacji o zróżnicowaniu ryzyka klientów. Te dwa modele binarne zostały użyte jako parametry *a priori* podejścia bayesowskiego. 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. Podejście bayesowskie pozwoliło na redukcję obciążenia modelu, które jest obserwowane w podejściu klasycznym.

Słowa kluczowe: małe próby, LGD, podejście bayesowskie, regresja logistyczna i liniowa, niezakończony proces odzysku