

Pitfalls in Modeling Loss Given Default of Bank Loans

by Marc Gürtler* and Martin Hibbeln**

* **Professor Dr. Marc Gürtler**
Braunschweig Institute of Technology
Department of Finance
Abt-Jerusalem-Str. 7
38106 Braunschweig
Germany
Phone: +49 531 391 2895
Fax: +49 531 391 2899
E-mail: marc.guertler@tu-bs.de

** **Dr. Martin Hibbeln**
Braunschweig Institute of Technology
Department of Finance
Abt-Jerusalem-Str. 7
38106 Braunschweig
Germany
Phone: +49 531 391 2898
Fax: +49 531 391 2899
E-mail: martin.hibbeln@tu-bs.de

Pitfalls in Modeling Loss Given Default of Bank Loans

Abstract

The parameter loss given default (LGD) of loans plays a crucial role for risk-based decision making of banks including risk-adjusted pricing. Depending on the quality of the estimation of LGDs, banks can gain significant competitive advantage. For bank loans, the estimation is usually based on discounted recovery cash flows, leading to workout LGDs. In this paper, we reveal several problems that may occur when modeling workout LGDs, leading to LGD estimates which are biased or have low explanatory power. Based on a portfolio of bank loans, we analyze these issues and derive recommendations for action in order to avoid these problems. Due to the restricted observation period of recovery cash flows the problem of length-biased sampling occurs, where long workout processes are underrepresented in the sample, leading to an underestimation of LGDs. Write-offs and recoveries are often driven by different influencing factors, which is ignored by the empirical literature on LGD modeling. We propose a two-step approach for modeling LGDs of non-defaulted loans which accounts for these differences leading to an improved explanatory power. In some situations banks are interested in forecasting absolute losses, suggesting to model recovery cash flows instead of LGDs. While both models have a similar performance in forecasting absolute losses, only LGD models are able to forecast relative losses. For LGDs of defaulted loans, the type of default and the length of the default period have high explanatory power, but estimates relying on these variables can lead to a significant underestimation of LGDs. We propose a model for defaulted loans which makes use of these influence factors and leads to consistent LGD estimates.

Keywords: Credit risk; Bank loans; Loss given default; Forecasting

JEL classification: G21; G28

1 Introduction

For the description of the risk of a loan, the most central parameters are the probability of default (PD) and the loss given default (LGD). While a decade ago, the focus of academic research and banking practice was mainly on the prediction of PDs, recently substantial effort has been put into modeling the LGD. One reason is the requirement of the Basel II / III framework, according to which banks have to provide own estimates of the LGD when using the advanced internal ratings-based (A-IRB) approach or the IRB approach for retail exposures. Besides the regulatory requirement, accurate predictions of LGDs are important for risk-based decision making, e.g. the risk-adjusted pricing of loans, economic capital calculations, and the pricing of asset backed securities or credit derivatives (cf. Jankowitsch et al., 2008). Consequently, banks using LGD models with high predictive power can generate competitive advantages whereas weak predictions can lead to adverse selection.

There exist different streams of LGD related literature. Literature dealing with the relation between PDs and LGDs include Frye (2000), Altman et al. (2005), and Acharya et al. (2007). LGD models that seek to estimate the distribution of LGDs for credit portfolio modeling are Renault and Scaillet (2004) and Calabrese and Zenga (2010). Furthermore, there are several empirical studies that analyze influencing factors of individual LGDs. While most of the literature consists of empirical studies for corporate bonds, a smaller fraction focuses on bank loans, whether retail or corporate, which is mainly due to limited data availability. A survey of empirical studies with a classification into bank and capital market data can be found in Grunert and Weber (2009).

There are some relevant differences between LGDs of corporate bonds and bank loans. First, LGDs of bank loans are typically lower than LGDs of corporate bonds. According to Schuermann (2006), this empirical finding is mainly a result of the (on average) higher seniority of loans and a better monitoring. Second, LGDs of corporate bonds are typically determined on the basis of market values resulting in “market LGDs” whereas the LGDs of bank loans are usually “workout LGDs”. If the market value of a bond directly after default is divided by the exposure at default (EAD), which is the face value at the default event, we get the market recovery rate (RR). Application of the equation $LGD = 1 - RR$ results in the market LGD. Contrary, the workout LGD is based on actual cash flows that are connected with the defaulted debt position. These are mainly discounted recovery cash flows but also discounted costs of the workout process. If these cash flows are divided by the EAD, we get the workout LGD. Even if the calculation of workout LGDs is more complex, the advantage

is that the results are more accurate and that this approach is applicable for all types of debt (cf. Calabrese and Zenga, 2010).

A first step towards forecasting individual LGDs of bank loans has been done by empirical studies reporting LGDs for different categories of influence factors (cf. Asarnow and Edwards, 1995; Felsovalyi and Hurt, 1998; Eales and Bosworth, 1998; Araten et al., 2004; Franks et al., 2004). More recent studies analyze influence factors of LGDs via linear regressions (cf. Citron et al., 2003; Caselli et al., 2008; Grunert and Weber, 2009), log regressions (cf. Caselli et al., 2008) or log-log regressions (cf. Dermine and Neto de Carvalho, 2005; Bastos, 2010). Belotti and Crook (2007) compare the performance of different models, constructed as combinations of different modeling algorithms and different transformations of the recovery rate, e.g. OLS regressions or decision trees on the one hand and log or probit transformations on the other hand. Bastos (2010) proposes to model LGDs with nonparametric and nonlinear regression trees.

The main motivation of this paper is to call attention to some pitfalls in modeling workout LGDs of bank loans. First, the empirical literature on LGDs widely ignores the effect that samples of historical LGDs are usually biased, which is due to differences in the length of the workout process. Two types of default end can be distinguished: contracts that can be recovered and contracts that have to be written off partly or completely. Since write-offs are typically connected with a longer period of the default status, the number of write-offs is usually underrepresented in samples of defaulted loans, leading to an underestimation of LGDs.

Second, due to the different characteristics of recovered loans and write-offs, it can be problematic to estimate LGDs with a single model. We propose a two-step estimation of LGDs: In the first step, the probability of a recovery/write-off is estimated. In the second step, the LGD of recovered loans as well as the LGD of write-offs is predicted separately. These predictions are combined into the total LGD forecast.

Third, although it is common to use the LGD or the RR, which are relative values, as target variable for predictions, it could also be argued that a prediction of the absolute loss is advantageous for a bank. Against this background, we compare both approaches. However, even if the explanatory power, at the first glance, seems to be higher when predicting absolute losses, we show that banks should focus on the estimation of the relative parameter LGD.

Fourth, not only for non-defaulted but also for defaulted loans with active default status, estimates of LGDs are required, e.g. for regulatory and economic capital calculations. Nevertheless, the existing literature on LGD modeling only concentrates on non-defaulted

loans. For defaulted loans, there is some additional information available that can be used for LGD predictions, e.g. we find that the length of the default period has a high explanatory power. However, if LGDs are modeled on the basis of the (ex post known) total length of default and the model is applied using the (ex ante known) current length of default, LGDs will be significantly underestimated. Thus, we show how the ex ante information of the current length of default can be used appropriately.

These aspects can significantly influence the forecasts and should be considered when modeling LGDs to achieve reasonable results. However, pitfalls 1 and 2 are mostly not taken into account in empirical studies, and, to our best knowledge, pitfalls 3 and 4 have not been addressed in the literature before. There are some further interesting findings. Within the first step of our estimation, i.e. the prediction of recovery/write-off probabilities, we find that the accuracy is lower for secured loans than for unsecured loans. However, within the second step, i.e. the prediction of LGDs conditional on the type of default end, the opposite is true. Furthermore, we propose a simple but well working model for estimating LGDs of defaulted loans.

The remainder of this paper is structured as follows. Section 2 contains a description of the data. In this context, we give attention to the first pitfall. In Section 3, we discuss LGD modeling for non-defaulted loans including pitfalls 2 and 3. Section 4 deals with LGD modeling for defaulted loans, which covers pitfall 4. Section 5 concludes.

2 Calculation of workout LGDs and description of the data set

For the forecasting of LGDs, we have to calculate historical workout LGDs of our modeling data. The workout LGD of loan i is typically expressed as follows:

$$LGD_i = 1 - \frac{\sum_{j=1}^J RCF_{i,j} - \sum_{k=1}^K C_{i,k}}{EAD_i}, \quad (1)$$

where RCF stands for the discounted recovery cash flows, C represents the discounted direct and indirect costs, and EAD is the exposure at default. However, a defaulted loan can have two different types of default end, which directly influence the calculation of LGDs: Some contracts can be recovered whereas other contracts have to be written off partly or completely. In the case of a recovery (RC), the default reason is no longer existent, e.g. the obligor paid the outstanding rates or a new payment plan is arranged. While equation (1) is correct for write-offs (WOs), we have to consider the exposure at recovery (EARC) for the case of RCs.

Since the EARC reduces the economic loss resulting from this default but the EARC is not included in the cash flows, we have to consider this value additionally:

$$LGD_i = 1 - \frac{\sum_{j=1}^J RCF_{i,j} - \sum_{k=1}^K C_{i,k} + EARC_i}{EAD_i}, \quad (2)$$

where we can set the value of EARC to zero if the type of default end is a write-off.

We apply equation (2) to calculate the LGDs of defaulted loans for a data set of a large German bank. The data set consists of 71,463 loans with default end between October 1st, 2006 and September 30th, 2008.¹ The loans correspond to several subportfolios of the bank, which can be divided into retail and commercial clients, secured and unsecured loans, as well as contracts with and without balloon payment. The LGD distribution corresponding to a subportfolio consisting of secured retail loans is presented in Figure 1.

- Figure 1 about here -

In the empirical literature about LGDs it is often reported, that the distribution of LGDs is bimodal with most LGDs being quite high (20-30%) or quite low (70-80%) (cf. Schuermann, 2006). While this seems to be true for corporate bonds or combined data of corporate bonds and corporate loans, the distribution for retail loans can be quite different. For our data set, it is striking that the major share of loans has a LGD which is close to zero, whereas a smaller share of loans is concentrated at values around 50%. This distribution has similarities to the data set of Bastos (2010). However, in our data the fraction of LGDs close to zero is considerably higher whereas the fraction of LGDs close to one is substantially lower. The large amount of contracts with LGDs close to zero mainly consists of loans that have been recovered. Observations with high LGDs largely belong to contracts that had to be written off. The distribution of LGDs for both types of default end, RC and WO, are illustrated in Figure 2.

- Figure 2 about here -

Banks are mainly interested in the total LGD of contracts and not only in the loss in a predefined period after default. For example, Bastos (2010) mentions for his study that the

¹ While most studies on LGDs present the number of loans that defaulted in a given period (default begin), we focus on the default end. Details will be described subsequently.

dates of write-offs were not available, but that LGDs calculated on the basis of recovery cash flows within a long time period after default are a good approximation of the demanded LGDs. Thus, if there is sufficient data available, only contracts with realized default end (RC or WO) should be considered in the modeling data. However, if we develop LGD models on the basis of all defaults with completed workout process that are available, defaults with a short workout process are overrepresented, which is due to interval censored data. This is illustrated in Figure 3.

- Figure 3 about here -

Since a short default length is mostly connected with small LGDs, the consequence of the overrepresentation of these data sets is an underestimation of LGDs. The impact of this underestimation is the greater, the shorter the time period that is covered by the data of a bank. The relevance of this issue becomes apparent if we look at the minimum data requirements for own estimates of LGDs according to the implementation of the regulatory capital rules (Basel II) into German law (Solvabilitätsverordnung, SolvV). According to § 133 and § 134(4) SolvV, LGD estimates must be based on a data observation period of at least 5 years for corporate and 2 years for retail exposures, if the bank uses own estimates of LGDs for the first time. Subsequently, the minimum data observation period increases to 7 and 5 years, respectively. For these data observation periods, the problem of uncompleted defaults can lead to a significant underestimation of LGDs.

Pitfall 1: Underestimation of LGDs due to restricted data observation periods

If we were solely interested in the duration of the workout process, we could account for censoring e.g. by using the proportional hazard or accelerated lifetime model.² However, we want to determine the LGDs of censored data and not the duration, so that we cannot apply these methods. In order to account for the bias of LGDs, we first analyze the distribution of the default length. Since recovered loans and write-offs have very different characteristics, we present the length of the default period for both types of default separately in Figure 4.

- Figure 4 about here -

² The estimation of the survival function for censored data using nonparametric and parametric methods is described in Kiefer (1988).

Typically, the workout process is significantly shorter for loans that can be recovered than for write-offs. Since recoveries usually have smaller LGDs than write-offs, as demonstrated before in Figure 2, we have an essential reason for the finding that defaults with a short default length typically have small LGDs. However, for the presented data almost all workout processes are completed after 450 days. Thus, if we do not consider all available default data but only those that could have been recovered or written off within 450 days, we avoid the systematical underestimation of LGDs. There are two ways of assuring this.

First, we can reduce the data set to loans with *default begin* between the beginning of the observation period and 450 days before the end of the observation period. Second, we can restrict the data to loans with *default end* between 450 days after the beginning of the observation period and the end of the observation period. We use the second alternative since only in this case we consider the most recent defaults and reject defaults from the beginning of the observation period. Contrary, if we chose the first alternative, we would have ignored the most recent defaults. Since our observation period comprises the time period between July 1st, 2005 and September 30th, 2008 we restrict the analysis to loans with default end between September 24th, 2006 and September 30th, 2008.

Even if this bias can lead to a significant underestimation of LGDs, in most empirical studies there is no remark that this potential bias is accounted for. For example, Grunert and Weber (2009) analyze loans which defaulted between 1992 and 2003. They note that only loans with completed workout process are considered, leading to a small number of defaults in the years 2002 and 2003. Thus, the mentioned bias has apparently not been accounted for. The same is true for Asarnow and Edwards (1995), even if the bias should be less substantial, which is due to the long data observation period from 1970 to 1993. As mentioned before, Bastos (2010) calculates LGDs on the basis of recovery cash flows within a recovery horizon of 12, 24, 36, and 48 months, where especially the recovery horizon of 48 months could be used as an approximation of the required LGD. Against this background, the author only considered defaults within the first 2 out of a 6 years data observation period. They thus do not consider the most recent defaults. The same is true for the empirical study of Dermine and de Carvalho (2006), where only the first 154 out of 374 defaults are considered for the recovery horizon of 48 months.

3 LGD forecasting for non-defaulted loans

3.1 Methodology of LGD modeling

Most of the empirical literature regarding influence factors of LGDs performs linear regressions and sometimes log or log-log-regressions with target variable LGD or RR. However, only a few studies have focused on forecasting LGDs (cf. Bastos, 2010). We find that the predictive power of the mentioned approaches is very low for our data set. When analyzing the data in detail, we have found that the characteristics of recovered loans are often very different from loans that have to be written-off. Thus, it seems reasonable to explicitly account for the differences between write-offs and recovered loans in the methodology of LGD forecasting.

Pitfall 2: Neglecting differences between write-offs and recovered loans in LGD forecasting

In order to account for the different characteristics of write-offs (WO) and recovered loans (RC), we estimate the LGDs with a two-step model. As a first step, we estimate the probability $\hat{\lambda}_{\text{WO}}$ of a write-off. Accordingly, the probability of a recovery is $\hat{\lambda}_{\text{RC}} = 1 - \hat{\lambda}_{\text{WO}}$. In the second step, we determine the LGDs for both types of default end separately, which leads to LGD forecasts $\widehat{LGD}_{\text{WO}}$ and $\widehat{LGD}_{\text{RC}}$. Finally, for each credit i , with $i = 1, \dots, n$, these estimates can be combined into an LGD forecast, which is given by

$$\widehat{LGD}_i = \hat{\lambda}_{\text{WO},i} \cdot \widehat{LGD}_{\text{WO},i} + (1 - \hat{\lambda}_{\text{WO},i}) \cdot \widehat{LGD}_{\text{RC},i}. \quad (3)$$

The probability of a write-off $\hat{\lambda}_{\text{WO}}$ is estimated using a logistic regression model:

$$E\left(1_{\{\text{WO}\},i} \mid x_{1,i}, \dots, x_{k,i}\right) = \hat{\lambda}_{\text{WO},i} = \frac{1}{1 + \exp(-z_i)} \quad \text{with} \quad z_i = \beta_0 + \sum_{j=1}^k \beta_j \cdot x_{j,i}, \quad (4)$$

where $1_{\{\text{WO}\},i}$ is an indicator variable, which equals one if credit i is written-off and zero otherwise. The variables $x_{1,i}, \dots, x_{k,i}$ correspond to k different characteristics, which can be borrower, loan or collateral specific. In cases where it is not possible to develop a model with sufficient predictive power, the probability $\hat{\lambda}_{\text{WO}}$ is set to the historical average write-off rate of the respective subportfolio.

In the second step, we perform linear regressions for estimating the LGD of loans that have to be written-off:

$$\widehat{LGD}_{\text{wo},i} = \gamma_0 + \sum_{j=1}^m \gamma_j \cdot y_{j,i}, \quad (5)$$

where $y_{1,i}, \dots, y_{m,i}$ are m different variables, which are borrower, loan or collateral specific. Since the LGDs of recovered loans, in contrast to write-offs, mostly have only small variations and these variations could not be predicted accurately, we assign the EAD-weighted historical average LGD for this type of default end:

$$\widehat{LGD}_{\text{RC},i} = \sum_{j=1}^N w_j \cdot LGD_{\text{RC},j}, \quad (6)$$

with $w_j := EAD_j / \sum_{n=1}^N EAD_n$. Our methodology is related to the modeling approach of Belotti and Crook (2007). In one of their models, Belotti and Crook (2007) apply the following two-step approach: In the first step, it is determined whether $LGD = 0$, $LGD = 1$, or $0 < LGD < 1$.³ In the second step, the case $0 < LGD < 1$ is modeled with linear regressions. However, in our setting we do not model the final outcome of the LGD but the recovery-/write-off-probability. Even if a recovery is often associated with very low outcomes of the LGD, the case $LGD = 0$ only coincides for a part of the data. Moreover, we did not find different characteristics for defaults with $LGD = 1$. Consequently, for our data set we get more reasonable results if the target variable is the type of default end (recovery or write-off).

The application of our two-step approach for exemplary subportfolios is presented subsequently.

3.2 Application of the two-step model

The models for estimating LGDs are developed with SAS[®] Enterprise Miner. The models for forecasting the write-off probabilities $\hat{\lambda}_{\text{wo}}$ are estimated using multivariate logit-regressions according to (4). Since the data base is sufficiently large, we do not use a k-fold cross-validation like Belotti and Crook (2007) or Bastos (2010) but split the data into 70% training data and 30% validation data. For many of the used categorical variables, the out-of-sample performance could be improved by aggregating the variables to a smaller number of classes, e.g. using the variables “limited liability” or “unlimited liability” instead of the concrete legal form of a company. The predictive power of the different logit-models is mainly evaluated with the receiver operating characteristic (ROC) for the validation data.⁴

³ The authors model recovery rates and not LGDs, but due to $LGD = 1 - RR$ this distinction does not matter.

⁴ Interestingly, when checking the economical plausibility, i.e. the concordance with the working hypotheses, the ROC curves for the training and the validation data generally become more similar if variables with

The ROC curve plots the “sensitivity”, i.e. the true positives, on the ordinate and “1 – specificity”, i.e. the false positives, on the abscissa. The ROC curves for the training and for the validation data, which correspond to the model of choice for one of the secured subportfolios, are presented in Figure 5. The respective values for the area under the ROC curve (AUC) are $AUC_{\text{Train}} = 73.5\%$ and $AUC_{\text{validate}} = 71.3\%$. As a final step, the coefficients of the model are calibrated on the basis of the full data set, leading to an AUC value of $AUC_{\text{All}} = 73.0\%$. The explanatory variables, which are used in the models, are borrower characteristics (e.g. the liability of a company for commercial clients or occupational category and marital status for private customers), collateral characteristics and loan characteristics (e.g. the previous number of defaults or the ratio of final balloon payment and residual value of the collateral). Interestingly, for unsecured loans it was possible to develop a model where the explanatory power is significantly higher, with $AUC_{\text{Train}} = 81.6\%$ and $AUC_{\text{validate}} = 82.2\%$ (cf. Figure 6).

- Figure 5 about here -

- Figure 6 about here -

Similarly, we develop the linear regression models for estimating LGDs in the scenario of a write-off. Thus, we split the data set of contracts which had to be written-off into training and validation data and perform multivariate linear regressions. The predictive power of the models is mainly evaluated with the coefficient of determination for the validation data R_{validate}^2 . This out-of-sample statistics is computed as

$$R_{\text{validate}}^2 = 1 - \frac{\sum_{i=1}^M (LGD_i - \widehat{LGD}_i)^2}{\sum_{i=1}^M (LGD_i - \overline{LGD}_{\text{Train}})^2}, \quad (7)$$

where $\overline{LGD}_{\text{Train}}$ is the historical average LGD of the training data, \widehat{LGD}_i (with $i = 1, \dots, M$) are the forecasted LGDs of the validation data (applying the model which is based on the training data), and LGD_i are the realized LGDs of the validation data.⁵ For secured loans to private customers, the coefficients of determination for the selected model are $R_{\text{Train}}^2 = 19.9\%$

implausible coefficients are dropped, resulting in a reduced performance for the training data but an increased predictive power for the validation data.

⁵ This out-of-sample R^2 statistic is also used in Campbell/Thompson (2008).

and $R_{\text{Validate}}^2 = 17.6\%$. The final coefficients are calibrated on the complete data set leading to $R_{\text{All}}^2 = 19.3\%$. Again, the explanatory variables can be classified into borrower characteristics (e.g. the occupational category for private customers), collateral characteristics (e.g. age, manufacturer, and residual value of a car), and loan characteristics (e.g. 1/EAD or down payment/EAD). Remarkably, when developing LGD models for unsecured loans to private customers, the predictive power of write-off LGDs was so low that the (exposure-weighted) average write-off LGD is assigned in this scenario. Thus, we find that for secured loans to private customers the accuracy when predicting write-off probabilities is lower than for unsecured loans, but within the second step, the prediction of LGDs in the case of write-offs, the opposite is true.

3.3 Target variable for predicting losses: relative or absolute values?

It has already been presented in the previous section, that the LGD is used as a target variable in the second step for predicting losses in the case of write-offs. This leads to identical results as choosing the target variable recovery rate defined as $RR = 1 - LGD$. However, it could also be reasonable to focus on recovery cash flows (RCFs) instead of LGDs. If we think about an OLS regression, choosing LGD as a target variable means that every credit has the same weight in the regression. This is because the sum of squared deviations between the predicted and the observed LGD is minimized, each of which are relative values. If our target variable is the RCF instead, we minimize the error in absolute instead of relative terms. Thus, the estimation is mainly influenced by large credits whereas small credits have a minor impact on the resulting functional form. On the one hand, it could be argued that a bank is mainly interested in the total amount of losses and that a good estimation for large credits is more important than the estimation for small credits. On the other hand, an estimation of cash flows can easily lead to large relative errors for small loans which directly result in a strongly inaccurate interest rate. The consequence of this misclassification could be an adverse selection for small loans, which could overweight the benefit of an improved estimation for larger loans.

Pitfall 3: Regression on recovery cash flows

In order to analyze the effect of estimating RCFs instead of LGDs, we implement both regressions and compare the results. As in the previous section, we split the data set into a training data set (70%) and a validation data set (30%). The coefficients of determination (R^2)

for the finally specified models with target variable LGD and RCF for both the training and the validation data sets are presented in Table 1. The out-of-sample R^2 statistic is defined as in (7). Thus, if the out-of-sample R^2 is positive, the forecast shows a better performance than the historical average.

- Table 1 about here -

At a first glance, the RCF regression performs significantly better than the LGD regression. However, the statistics above are not appropriate for comparing the performance of the models. For a appropriate comparison, LGD predictions have to be transformed into RCF predictions (I) or RCF predictions have to be transformed into LGD predictions (II). Only then we are able to evaluate the performance of both model types.

$$(I): \widehat{RCF}_i = \widehat{RR}_i \cdot EAD_i = (1 - \widehat{LGD}_i) \cdot EAD_i, \quad (8)$$

$$(II): \widehat{LGD}_i = 1 - \widehat{RR}_i = 1 - \frac{\widehat{RCF}_i}{EAD_i}. \quad (9)$$

The effect, that a small predictive power of an LGD model can have a high predictive power for RCFs, can best be seen when we use the historical average LGD for predicting LGDs. Naturally, the coefficient of determination equals zero for the training data, and we could expect the R^2 for the validation data to be close to zero since we do not have a problem with overfitting the data. However, when we apply equation (8) to transform the LGD prediction into an RCF prediction, we get quite high coefficients of determination (see Table 2). The reason is, that the EAD of a loan already explains recovery cash flows to a large extent, whereas dividing by EAD for calculation of the LGD cancels out this explanatory variable.

- Table 2 about here -

Against this background, we transform the results of our LGD and our RCF model from Table 1 using equation (8) and (9) (see Table 3). The results show that the LGD model performs almost as good in predicting recovery cash flows as the RCF model. This shows that a direct comparison of the R^2 from Table 1 is misleading. Interestingly, the RCF model leads to good predictions of recovery cash flows but miserably fails in predicting LGDs. Especially, the negative values of R^2 show that prediction of LGDs using the RCF model lead to an even worse performance than the historical average LGD. This approves that a RCF model is very problematic for assigning appropriate interest rates to smaller loans. Consequently, we find

that whether a bank is mainly interested in a good forecasting of relative or absolute losses, it is highly recommendable to develop a model for loss given default instead of recovery cash flows.⁶

- Table 3 about here -

4 LGD forecasting for defaulted loans

For defaulted loans, the parameters PD and EAD are realized values but the LGD is still a random variable. However, we have some additional information about the loan which can be used for LGD forecasting. Especially, we have knowledge about the default reason and the current length of the default period:

- The concrete events which characterize the default of a loan vary from bank to bank. Some typical reasons are (1) the obligor is past due for more than 90 days, (2) a notice of cancellation, (3) a court order, or (4) a significant downgrading. We find that the average LGD varies significantly depending on different default reasons. For example, defaults with default reason 1 (being past due) on average lead to smaller losses than defaults with default reason 2 (notice of cancellation).
- Furthermore, the average LGD of contracts with a long default period is usually higher than the LGD of contracts with a short default period. A part of this effect stems from the on average different default periods of loans that can be recovered and loans that have to be written off (cf. section 2). Additionally, even within the write-offs, the LGDs are mostly higher for contracts with a long default period.

In order to analyze which factors are most important for explaining the LGD of defaulted loans, we use regression trees with the software SAS[®] Enterprise Miner.⁷ Regression trees are a nonlinear and nonparametric predictive modeling tool, which splits the data into several groups on the basis of a series of binary questions, e.g. “default reason = 1?” and “default period > 100 days?”. These questions are set in a way that the information about the LGD is

⁶ However, if analyses show that the specified model is very different for small and large loans, it should be considered to split the data set or to introduce dummy variables for different exposure classes. Moreover, for revolving retail credits the out-of-sample results can usually be improved when rejecting observations with EADs below a predefined threshold.

⁷ According to Bastos (2010), the first published study which models LGDs with regression trees is Bastos (2010). However, we apply regression trees to forecast LGDs of defaulted instead of non-defaulted loans.

maximized.⁸ As noticed by Bastos (2010), regression trees are well-suited for producing accurate results of LGD forecasts using only a few important explanatory variables. We find for different subportfolios that the most important explanatory variables are the default reason, the length of the default period and some segmentation variables regarding the type of obligor, loan, and collateral. However, we have to consider the different set of information about the default length of contracts with active and completed workout process. For modeling purposes, we have knowledge of the total length of default. Contrary, when applying the model to active defaults, we only know the current length of default, which is obviously smaller than the total length. Since the LGD is generally small when the default period is short, ignoring the difference between the information sets would lead to a significant underestimation of the LGD.

Pitfall 4: Underestimation of LGDs when using the total length of the default period as explanatory variable

Since the length of the default period has a high explanatory power for LGDs, we intend to use the known information set about the default length. We first partition our modeling data into classes which are homogeneous regarding the mentioned segmentation variables and the default reason. Within these classes, we specify LGDs regarding the “minimum default length” (MDL) because the current length of default is smaller or equal to the total default length (TDL). For this purpose, we calculate EAD-weighted average LGDs belonging to each value of MDL:

$$\begin{aligned}
 \widehat{LGD}_{\text{Default},t} &= E(LGD_i | TDL_i \geq t) \\
 &= \frac{\sum_{j=1}^N w_j \cdot LGD_j \cdot I\{TDL_j \geq t\}}{\sum_{j=1}^N w_j \cdot I\{TDL_j \geq t\}} \\
 &= \frac{\sum_{j=1}^N EAD_j \cdot LGD_j \cdot I\{TDL_j \geq t\}}{\sum_{j=1}^N EAD_j \cdot I\{TDL_j \geq t\}} =: LGD(MDL = t),
 \end{aligned} \tag{10}$$

where $j = 1, \dots, N$ stands for all contracts of our modeling data within a class and $I\{TDL \geq t\}$ takes the value one if the argument is true and zero otherwise. Since the information set of MDL of the modeling data is identical to the current default length of the scoring data, we get

⁸ For details see Breiman (1984).

consistent LGDs when we apply the model. However, for large values of MDL, we set the LGD to a constant value in order to reduce the estimation error resulting from the small number of observations. Moreover, since the empirical LGDs exhibit some economically implausible jumps or non-monotonous sections, we describe the rest of the function piecewise with polynomial functions. Graphical illustrations of the empirical LGDs resulting from equation (10), which correspond to one of the segments, are presented in Figure 7.

- Figure 7 about here -

There are some characteristics of the illustrations worth mentioning. First, default reasons 2 and 3 are aggregated since one of these categories is usually almost empty depending on whether the collateral has already been liquidated in a previous default or not.⁹ Second, for most contracts with default reason 1, 2, or 3, the LGD increases with the default length. Third, the average LGD of contracts with default reason 4 decreases for small values of MDL and has a jump at $t = 365$ days. To understand this effect, we have to consider that default reason 4 means a significant downgrading. Banks often retrieve additional scoring information from credit agencies. In the presented case of retail loans, the values of the negative scoring characteristics are updated one year after default. If the negative scoring characteristic is no longer existent and if this is the only active default reason at this time, a loan recovers, leading to a small LGD. This effect was already visible in Figure 4, where we could observe a small peak of recovered loans for a default length of 365 days. However, if default reason 4 is still existent, the probability of a write-off is quite high. Thus, the LGD has a jump at a minimum default length of one year.

5 Conclusion

In this paper, we identify some pitfalls in modeling workout LGDs which can easily lead to inaccurate LGD forecasts. First, the LGDs within the modeling data are biased downwards if all available defaults with completed workout process are considered. This is mainly due to the different default length of recovered loans and write-offs. We suggest considering only loans with default end within a predefined time period. Thus, we can use all recent data and only reject data with defaults in the beginning of the observation period. Second, we propose a two-step approach for modeling LGDs of non-defaulted loans. With this approach, we have

⁹ During the default period, the default status can change, e.g. from 2 to 3. However, the default reason remains unchanged.

achieved better predictions than with other approaches proposed in the literature, since different influencing factors of recoveries and write-offs can be considered. Third, we test whether it is advantageous to base LGD forecasts on absolute values (predicted recovery cash flows) or to directly focus on the relative variable LGD. We find that forecasts with target variable LGD have similar predictive power for forecasting absolute losses but are strictly preferable regarding relative losses. Consequently, in order to avoid adverse selection, it is advantageous to rely on forecasts using relative values. Fourth, we propose a simple model to forecast LGDs of defaulted loans. We find that both the type of default end and the default length have a high explanatory power when forecasting those LGDs. Since the actual default length of scoring data and the total default length of the modeling data include different information sets of the default length, the LGDs are significantly underestimated when this difference is neglected. However, a minimum default length can be constructed for the modeling data, which contains the same information set as the current default length of the scoring data, leading to consistent LGD estimates.

While some of our findings are generally valid, others could be specific for the data used. For example, we find that the predictive power for estimating the probability of a recovery or a write-off is higher for unsecured than for secured loans. However, conditional on the type of default end the opposite is true. Moreover, while we mainly focused on retail loans, our models could also be beneficial for corporate loans. This is left for further research.

References

- Acharya, V.V., Bharath, S.T., Srinivasan, A., 2007. Does industry wide distress affect defaulted firms? Evidence from creditor recoveries. *Journal of Financial Economics* 85, 787–821.
- Altman, E.I., Brady, B., Resti, A., Sironi, A., 2005. The link between default and recovery rates: Theory, empirical evidence, and implications. *Journal of Business* 78, 2203–2228.
- Araten, M., Jacobs Jr., M., Varshney, P., 2004. Measuring LGD on commercial loans: An 18-year internal study. *The RMA Journal* 4, 96–103.
- Asarnow, E., Edwards, D., 1995. Measuring loss on defaulted bank loans. A 24-year-study. *Journal of Commercial Lending*, Edition 77 (7), 11–23.
- Bastos, J.A., 2010. Forecasting bank loans loss-given-default. *Journal of Banking and Finance* 34, 2510–2517.
- Bellotti, T., Crook, J., 2007. Modelling and predicting loss given default for credit cards. Working paper, Quantitative Financial Risk Management Centre.

- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth: Belmont, CA.
- Calabrese, R., Zenga, M., 2010. Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking and Finance* 34, 903–911.
- Campbell, J.Y., Thompson, S.B., 2008. Predicting Excess Stock Returns Outof Sample: Can Anything Beat the Historical Average? *Review of Financial Studies* 21, 1509–1531.
- Caselli, S., Gatti, S., Querci, F., 2008. The sensitivity of the loss given default rate to systematic risk: new empirical evidence on bank loans. *Journal of Financial Services Research* 34, 1–34.
- Citron, D., Wright, M., Ball, R., Rippington, F., 2003. Secured Creditor Recovery Rates from Management Buy-Outs in Distress. *European Financial Management* 9, 141–161.
- Dermine, J., Neto de Carvalho, C., 2006. Bank loan losses-given-default: a case study. *Journal of Banking and Finance* 30, 1243–1291.
- Eales, R., Bosworth, E., 1998. Severity of loss in the event of default in small business and larger consumer loans. *The Journal of Lending and Credit Risk Management*, 58–65.
- Felsovalyi, A., Hurt, L., 1998. Measuring loss on Latin American defaulted bank loans: A 27-year study of 27 countries. *Journal of Lending and Credit Risk Management*.
- Franks, J., de Servigny, A., Davydenko, S., 2004. A comparative analysis of the recovery process and recovery rates for private companies in the UK, France, and Germany. Standard and Poor's Risk Solutions, June 2004.
- Frye, J., 2000. Collateral Damage. *Risk* 13(4), 91–94.
- Grunert, J., Weber, M., 2009. Recovery rates of commercial lending: empirical evidence for German companies. *Journal of Banking and Finance* 33, 505–513.
- Jankowitsch, R., Pullirsch, R., Veža, T., 2008. The delivery option in credit default swaps. *Journal of Banking and Finance* 32, 1269–1285.
- Kiefer, N.M., 1988. Economic duration data and hazard functions.
- Renault, O., Scaillet, O., 2004. On the way to recovery: A nonparametric bias-free estimation of recovery rates densities. *Journal of Banking and Finance* 28, 2915–2931.
- Schuermann, T., 2006. What Do We Know About Loss Given Default? In: Shimko, D. (Ed.), *Credit Risk Models and Management*, 2nd Edition. Risk Books: London.

Figure 1

Distribution of loss given default of secured retail loans

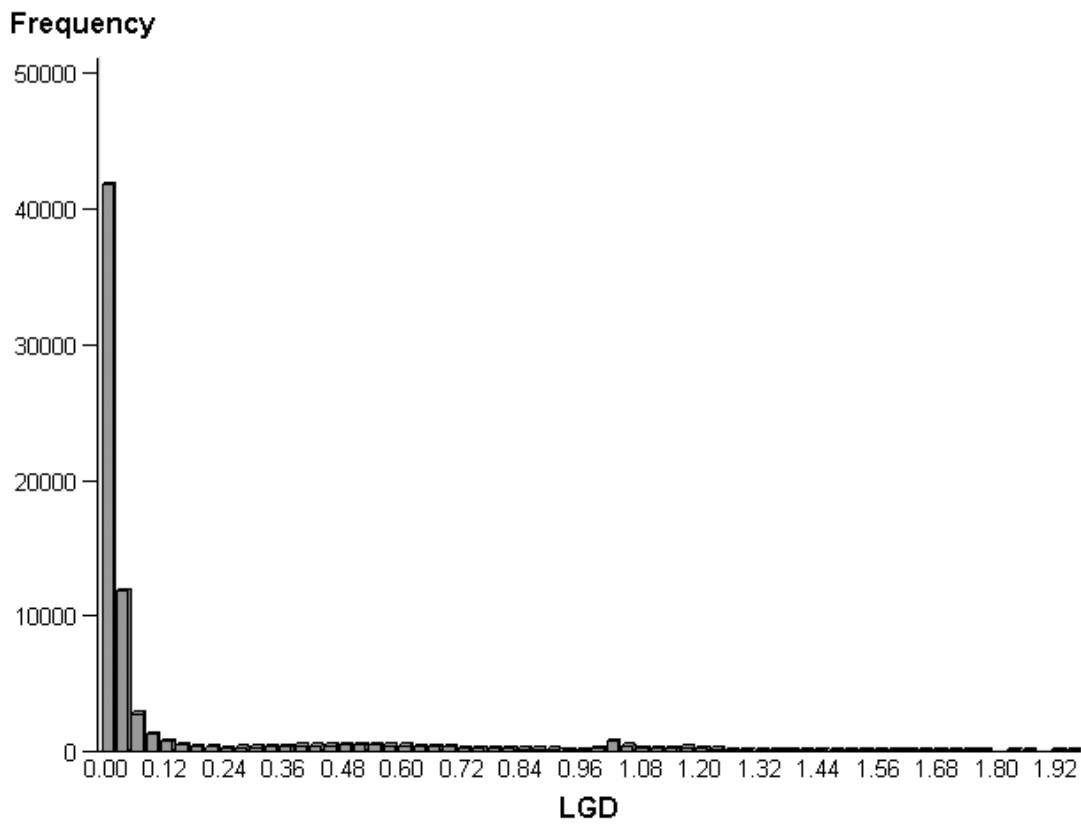


Figure 2

Distribution of loss given default for recovered loans (top) and for write-offs (bottom)

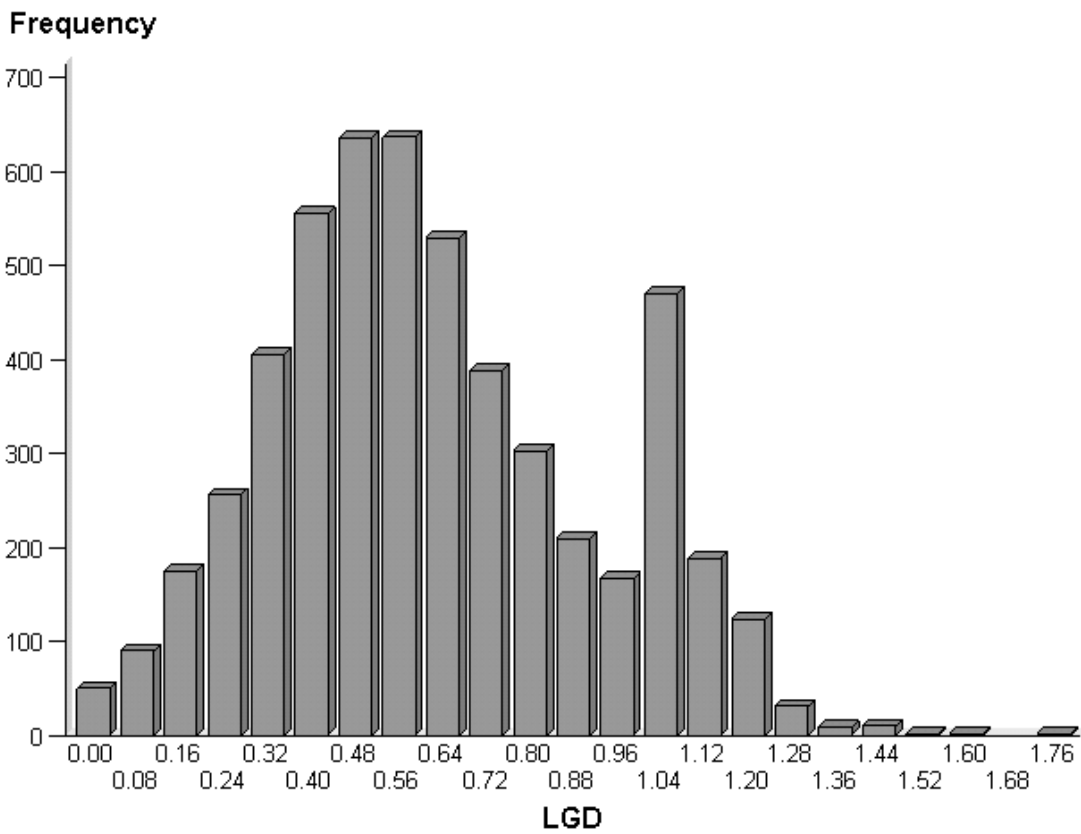
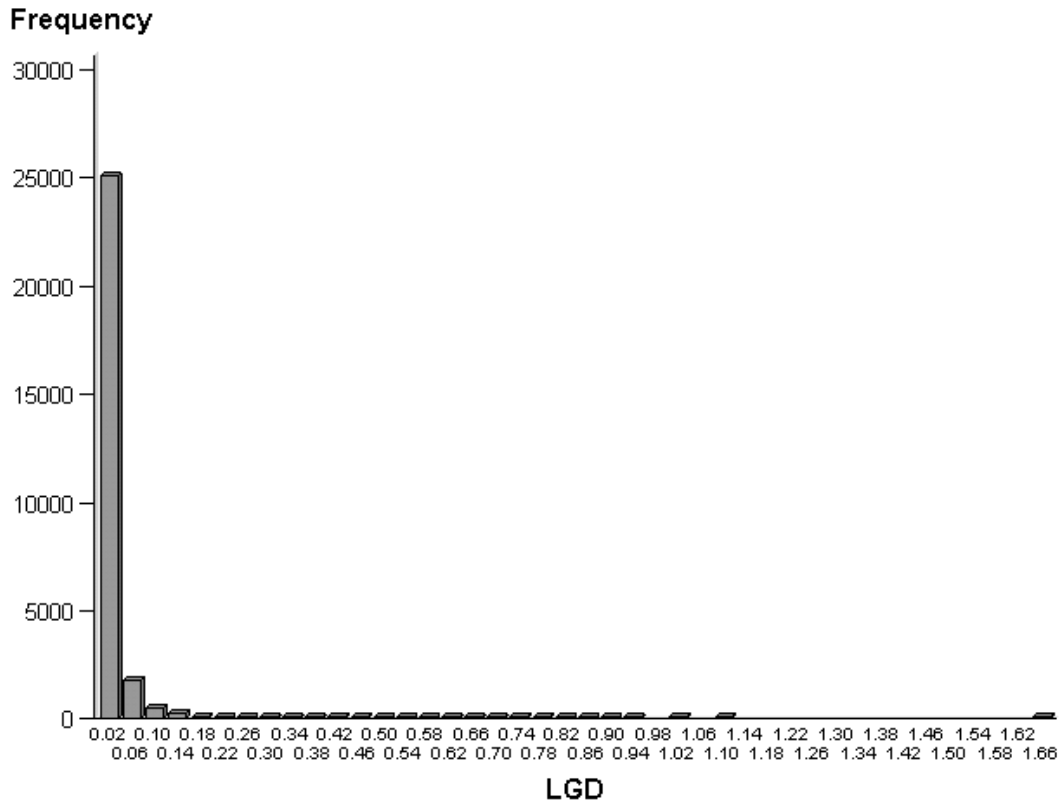


Figure 3

Interval censored data: Defaults with default begin and default end within the data observation period (completed workout process) are available in the data base (solid lines), other defaults are not included in the data base (dashed lines)

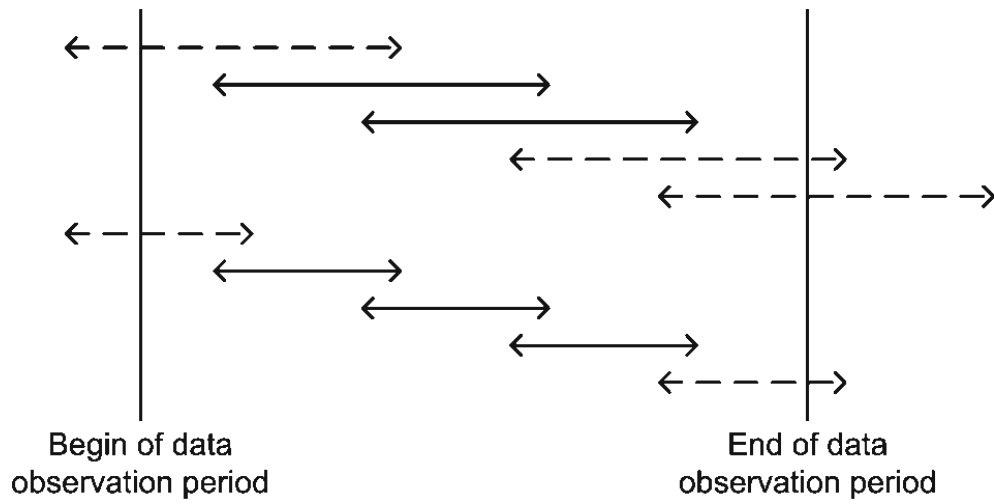


Figure 4

Length of the default period for recovered loans (top) and for write-offs (bottom) in days

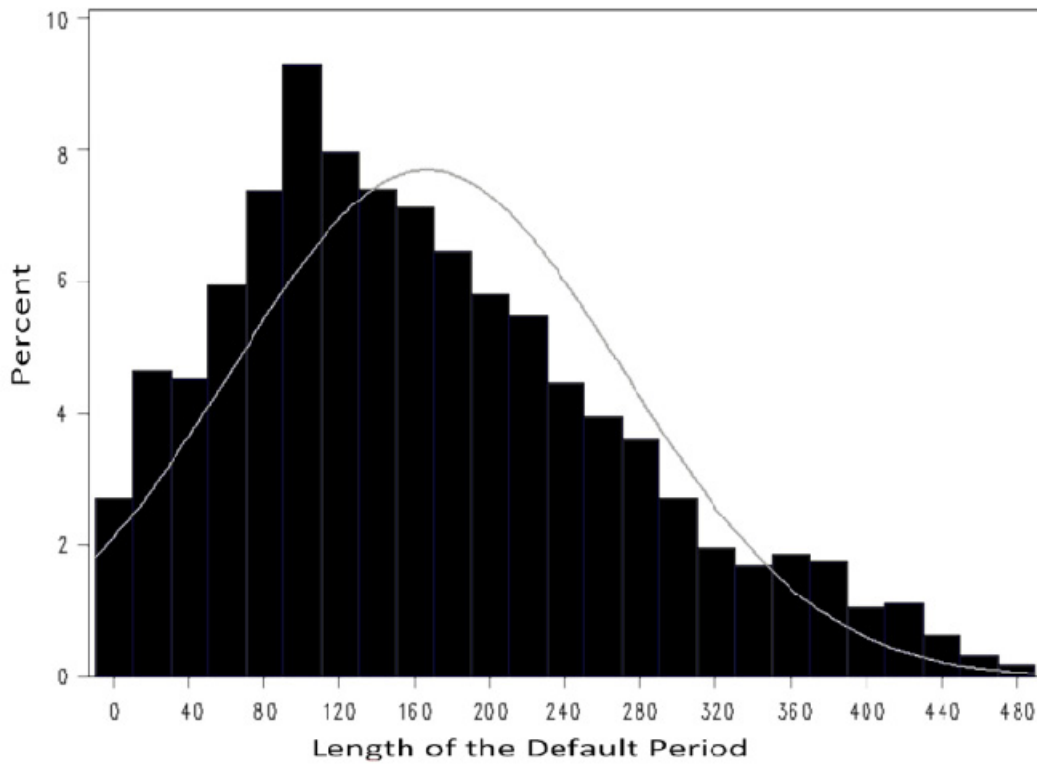
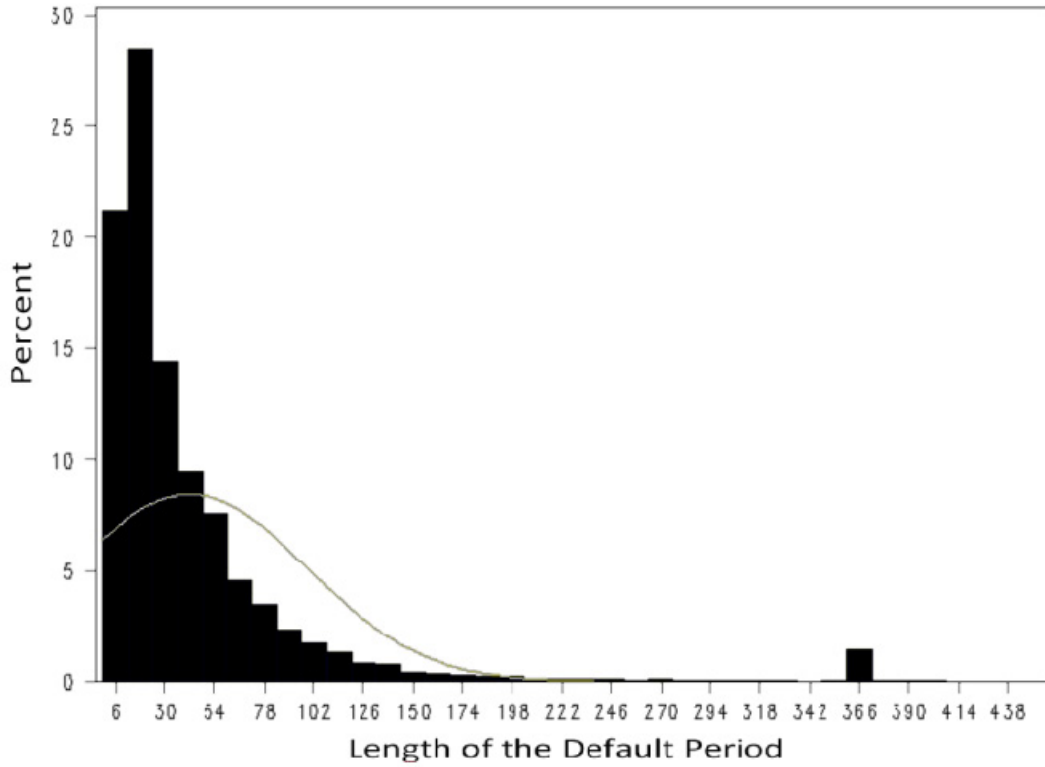


Figure 5

Receiver operating characteristic when forecasting write-off probabilities for the training (left) and validation data (right) of a secured subportfolio

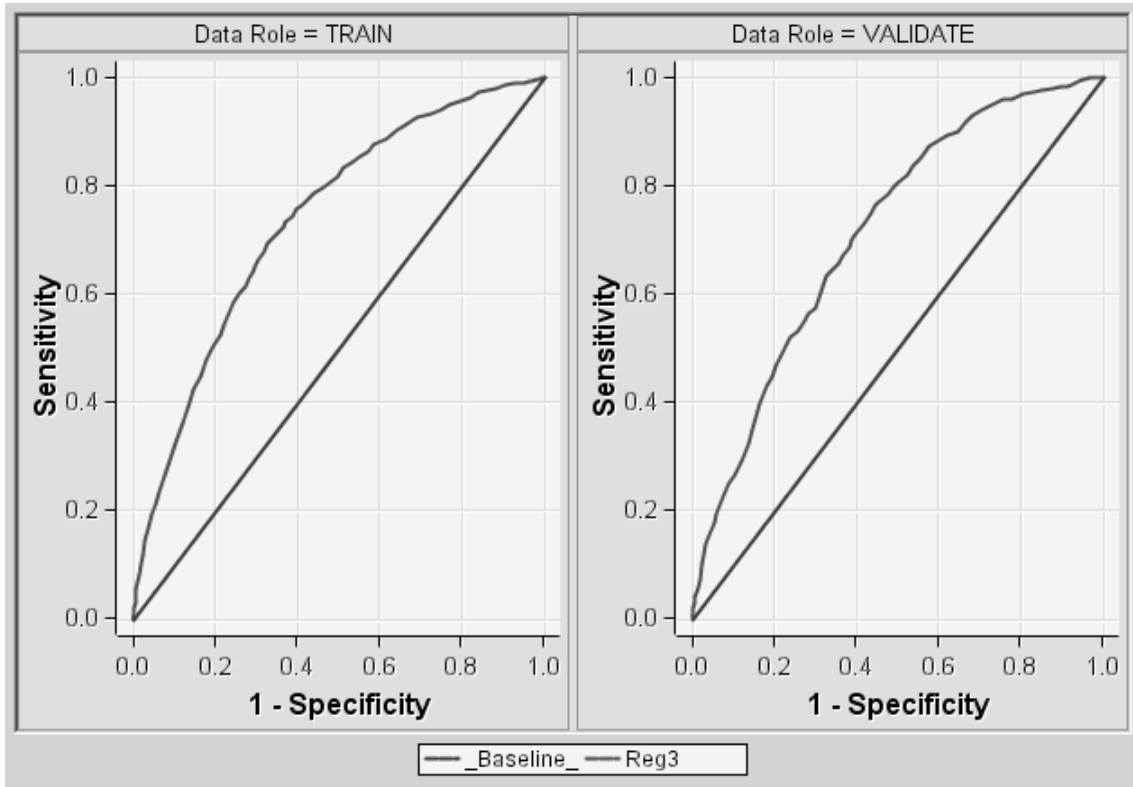


Figure 6

Receiver operating characteristic when forecasting write-off probabilities for the training (left) and validation data (right) of an unsecured subportfolio

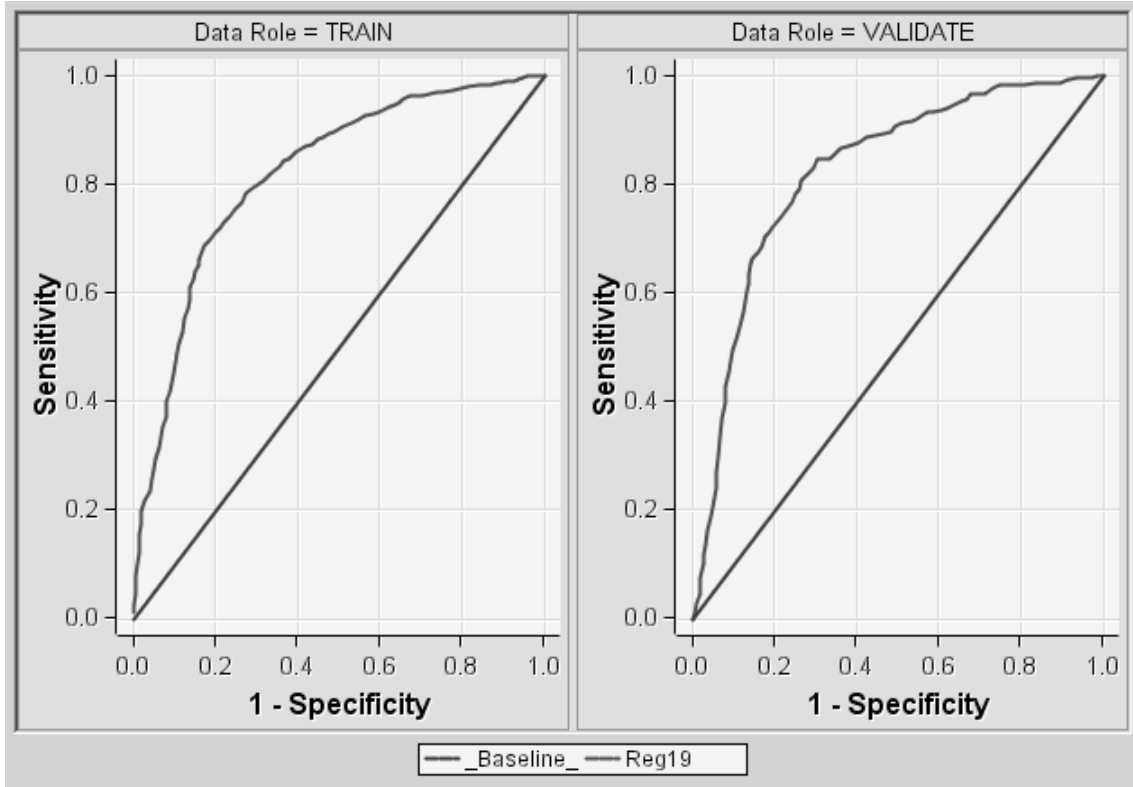


Figure 7

EAD-weighted LGDs (diamonds) and number of contracts (solid line) for default reason 1: being past due (top), default reason 2 & 3: notice of cancellation & court order (middle), and default reason 4: significant downgrading (bottom) depending on the minimum default length (in days)

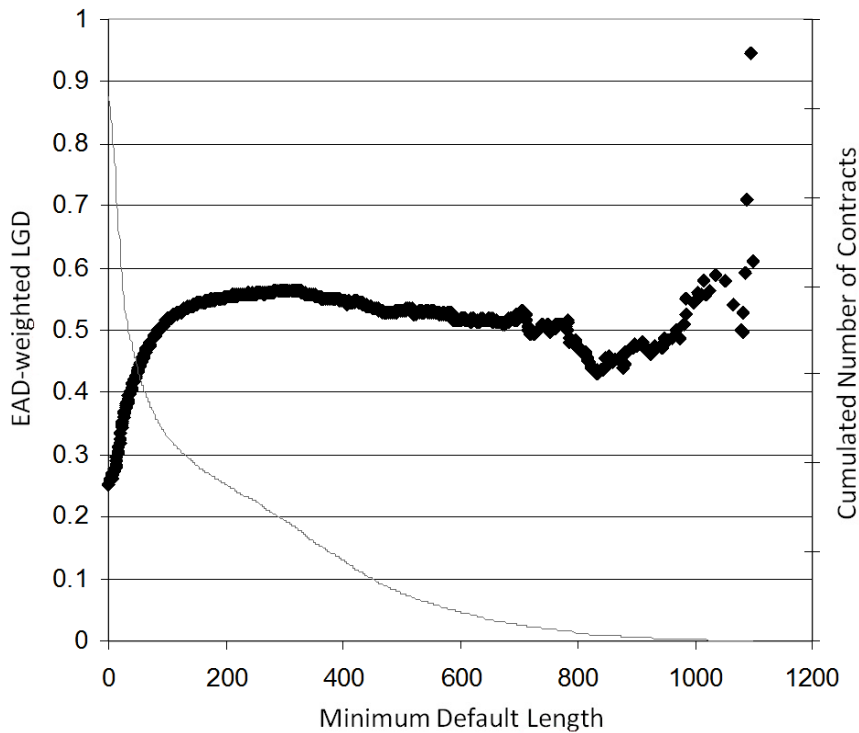
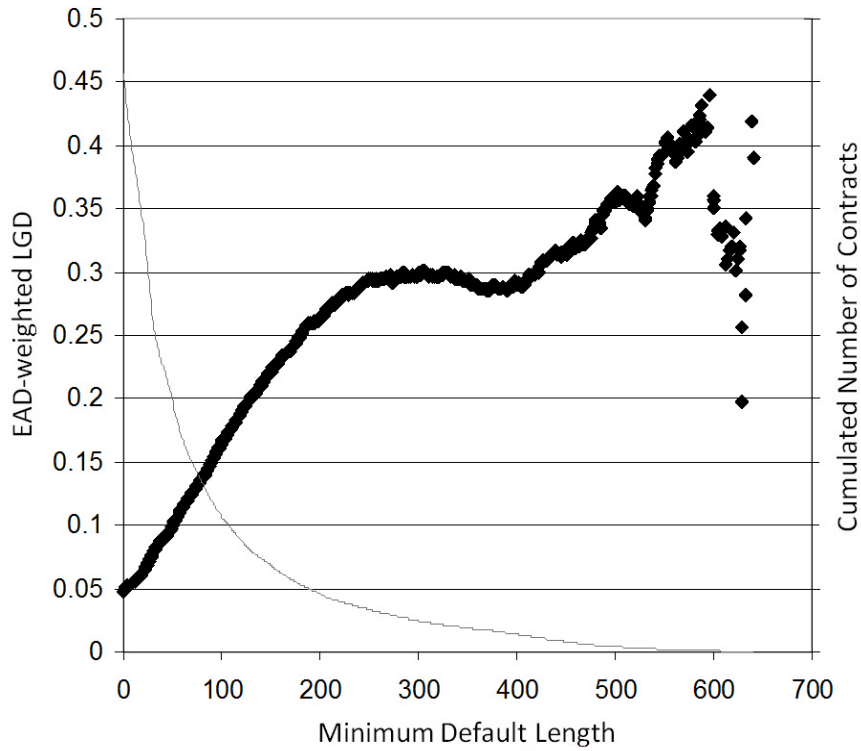


Figure 7 (continued)

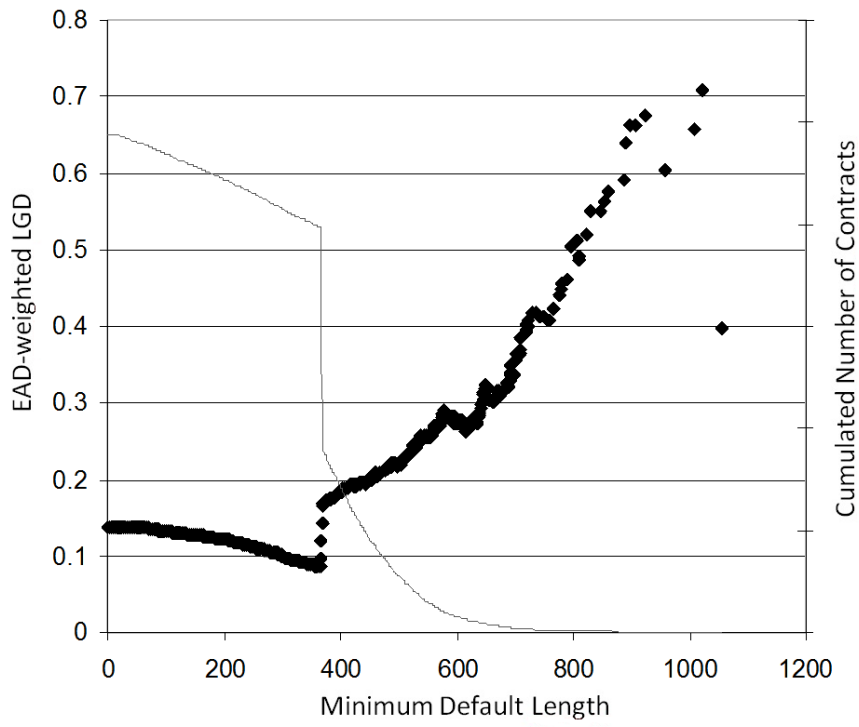


Table 1

R^2 for an LGD model and an RCF model on training and validation data

	R^2 of LGD model	R^2 of RCF model
Training data	19.89%	55.33%
Validation data	17.55%	57.43%

Table 2

R^2 of the historical average LGD used for LGD prediction (I) and RCF prediction (II) on training and validation data

	(I) R^2 of the historical average LGD for prediction of LGDs	(II) R^2 of the historical average LGD for prediction of RCFs
Training data	0%	48.17%
Validation data	0%	49.66%

Table 3

R^2 for an LGD model and an RCF model, each applied for LGD and RCF prediction, on training and validation data

	R^2 for prediction of LGDs		R^2 for prediction of RCFs	
	LGD model	RCF model	LGD model	RCF model
	Training data	19.89%	-1211.88%	52.23%
Validation data	17.55%	-810.47%	57.27%	57.43%